Quality Estimation for Language Output Applications

Carolina Scarton, Gustavo Paetzold and Lucia Specia

University of Sheffield, UK



COLING, Osaka, 11 Dec 2016

Quality Estimation

- Approaches to **predict** the quality of a language output application – no access to "true" output for comparison
- Motivations:
  - Evaluation of language output applications is hard: **no single gold-standard**
  - For NLP **systems in use**, gold-standards are not available
- Some work done for other NLP tasks, e.g. parsing

Quality Estimation - Parsing

Task [Ravi et al., 2008]

- Given: a statistical parser and its training data and some chunk of text
- Estimate of the f-measure of the parse trees produced for that chunk of text

Features

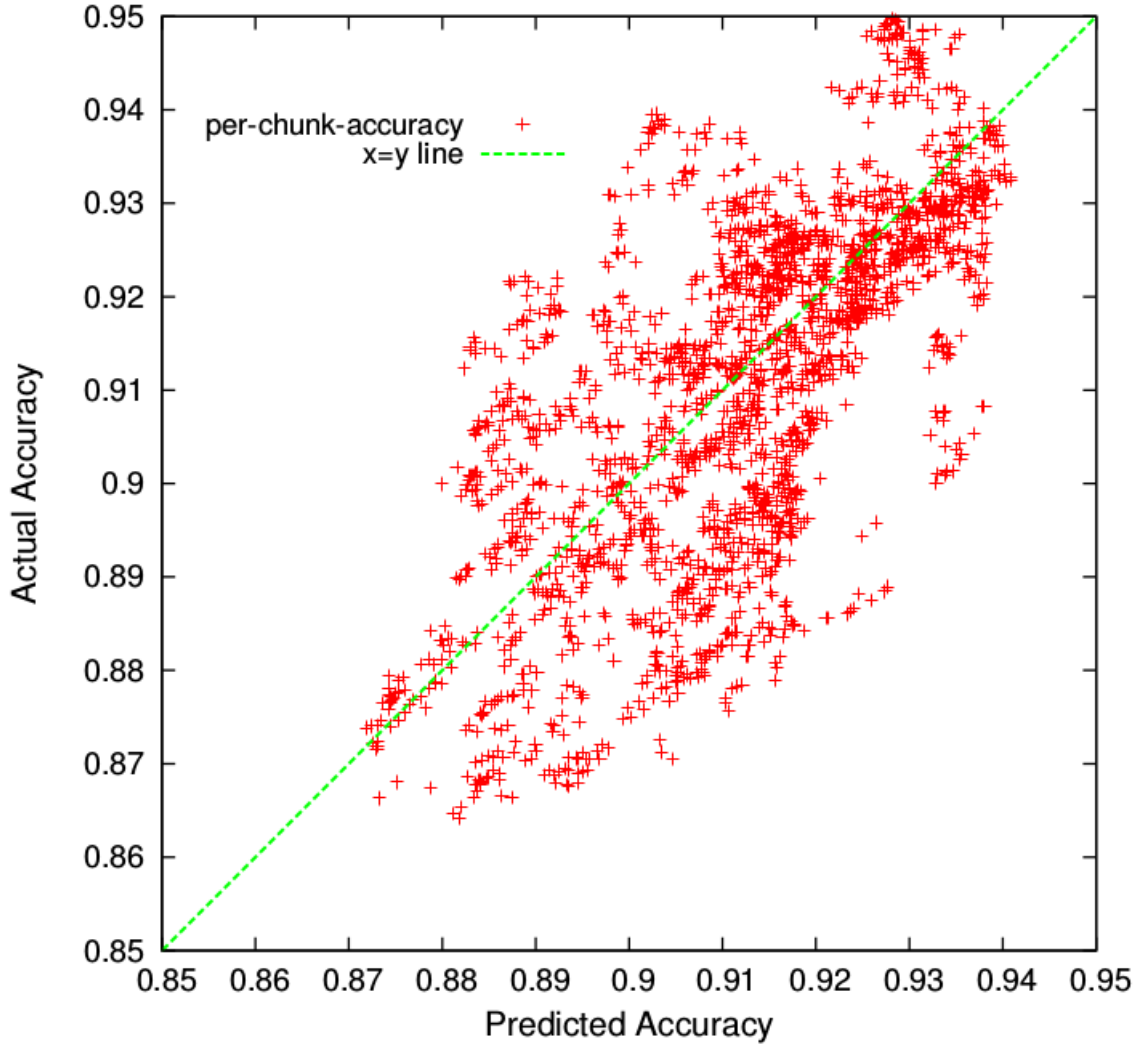
- Text-based, e.g. length, LM perplexity
- Parse tree, e.g. number of certain syntactic labels such as punctuation
- Pseudo-ref parse tree: similarity to output of another parser

Training

- Training data labelled for f-measure based on gold-standard
- Learner optimised for correlation of f-measure

Quality Estimation - Parsing

Very high correlation and low error (in-domain): RMSE = 0.014



Quality Estimation - Parsing

- Very close to actual f-measure:

	In-domain (WSJ)
Baseline (mean of dev set)	90.48
Prediction	90.85
Actual f-measure	91.13
	Out-of-domain (Brown)
Baseline (mean of dev set)	90.48
Prediction	86.96
Actual f-measure	86.34

- **Simpler task**: one possible good output; f-measure is very telling

Quality Estimation - Summarisation

Task: Predict quality of automatically produced summaries without human summaries as references [Louis and Nenkova, 2013]

- Features:
  - **Distribution similarity** and **topic words** → high correlation with PYRAMID and RESPONSIVENESS
  - **Pseudo-references**:
    - Outputs of off-the-shelf AS systems → additional summary models
    - High correlation with human scores, even on their own
  - Linear combination of features → **regression task**

Quality Estimation - Summarisation

[Singh and Jin, 2016]:

- Features addressing **informativeness** (IDF, concreteness, n-gram similarities), **coherence** (LSA) and **topics** (LDA)
- **Pairwise classification** and **regression** tasks predicting RESPONSIVENESS and linguistic quality
- Best results for **regression models** $\rightarrow$ RESPONSIVENESS (around 60% of accuracy)

Quality Estimation - Simplification

Task: Predict the quality of automatically simplified versions of text

- **Quality** features:
  - **Length** measures
  - Token **counts**/**ratios**
  - **Language model** probabilities
  - **Translation** probabilities
- **Simplicity** ones:
  - **Linguistic** relationships
  - **Simplicity** measures
  - **Readability** metrics
  - **Psycholinguistic** features
- **Embeddings** features

Quality Estimation - Simplification

QATS 2016 shared task

- The **first** QE task for Text Simplification

- **9** teams
- **24** systems

- **Training set:** 505 instances
- **Test set:** 126 instances

Quality Estimation - Simplification

QATS 2016 shared task

- 2 tracks:
  - **Regression:** 1/2/3/4/5
  - **Classification:** Good/Ok/Bad
- 4 aspects:
  - **Grammaticality**
  - **Meaning Preservation**
  - **Simplicity**
  - **Overall**

Quality Estimation - Simplification

QATS 2016 shared task: baselines

- Regression **and** Classification:
  - BLEU
  - TER
  - WER
  - METEOR
- Classification **only**:
  - Majority class
  - SVM with **all metrics**

Quality Estimation - Simplification

Systems:

System	ML	Features
UoLGP	**GPs**	**QuEst features** and **embeddings**
OSVCML	**Forests**	**embeddings**, **readability**, **sentiment**, etc
SimpleNets	**LSTMs**	**embeddings**
IIT	**Bagging**	**language models**, **METEOR** and **complexity**
CLaC	**Forests**	**language models**, **embeddings**, **length**, **frequency**, etc
Deep(Indi)Bow	**MLPs**	**bag-of-words**
SMH	**Misc.**	**QuEst features** and **MT metrics**
MS	**Misc**	**MT metrics**
UoW	**SVM**	**QuEst features**, **semantic similarity** and **simplicity metrics**

Quality Estimation - Simplification

Evaluation metrics:

- **Regression**: Pearson
- **Classification**: Accuracy

Winners:

	Regression	**Classification**
Grammaticality	OSVCML1	Majority-class
Meaning	IIT-Meteor	SMH-Logistic
Simplicity	OSVCML1	SMH-RandForest-b
Overall	OSVCML2	SimpleNets-RNN2

Quality Estimation - Machine Translation

Task: Predict the quality of an MT system output without reference translations

- **Quality**: fluency, adequacy, post-editing effort, etc.
- **General method**: supervised ML from features + quality labels

- Started circa 2001 - **Confidence Estimation**
  - How **confident** MT system is in a translation
  - Mostly word-level prediction from SMT internal features

- **Now**: broader area, commercial interest

Motivation - post-editing

MT: The King closed hearings Monday with Deputy Canary Coalition Ana Maria Oramas González -Moro, who said, in line with the above, that “there is room to have government in the coming months,” although he did not disclose prints Rey about reports Francesco Manetto. Monarch Oramas transmitted to his conviction that ‘ soon there will be an election” because looks unlikely that Rajoy or Sanchez can form a government.

Motivation - post-editing

MT: The King closed hearings Monday with Deputy Canary Coalition Ana Maria Oramas González -Moro, who said, in line with the above, that “there is room to have government in the coming months,” although he did not disclose prints Rey about reports Francesco Manetto. Monarch Oramas transmitted to his conviction that ‘ soon there will be an election” because looks unlikely that Rajoy or Sanchez can form a government.

SRC: El Rey cerró las audiencias del lunes con la diputada de Coalición Canaria Ana María Oramas González-Moro, quien aseguró, en la línea de los anteriores, que “no hay ambiente de tener Gobierno en los próximos meses”, aunque no desveló las impresiones del Rey al respecto, informa Francesco Manetto. Oramas transmitió al Monarca su convicción de que “pronto habrá un proceso electoral”, porque ve poco probable que Rajoy o Sánchez puedan formar Gobierno.

By Google Translate

Motivation - gisting

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

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

By Google Translate

Motivation - gisting

Target:
the road boycotted a friend ... indian robin hood killed the poor after 32 years of prosecution.

Source:
مقتل روبن هود الهندي.. قاطع الطريق صديق الفقراء بعد 32 عاما من الملاحقة

Reference:
death of the indian robin hood, highway robber and friend of the poor, after 32 years on the run.

By Google Translate

Uses

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

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

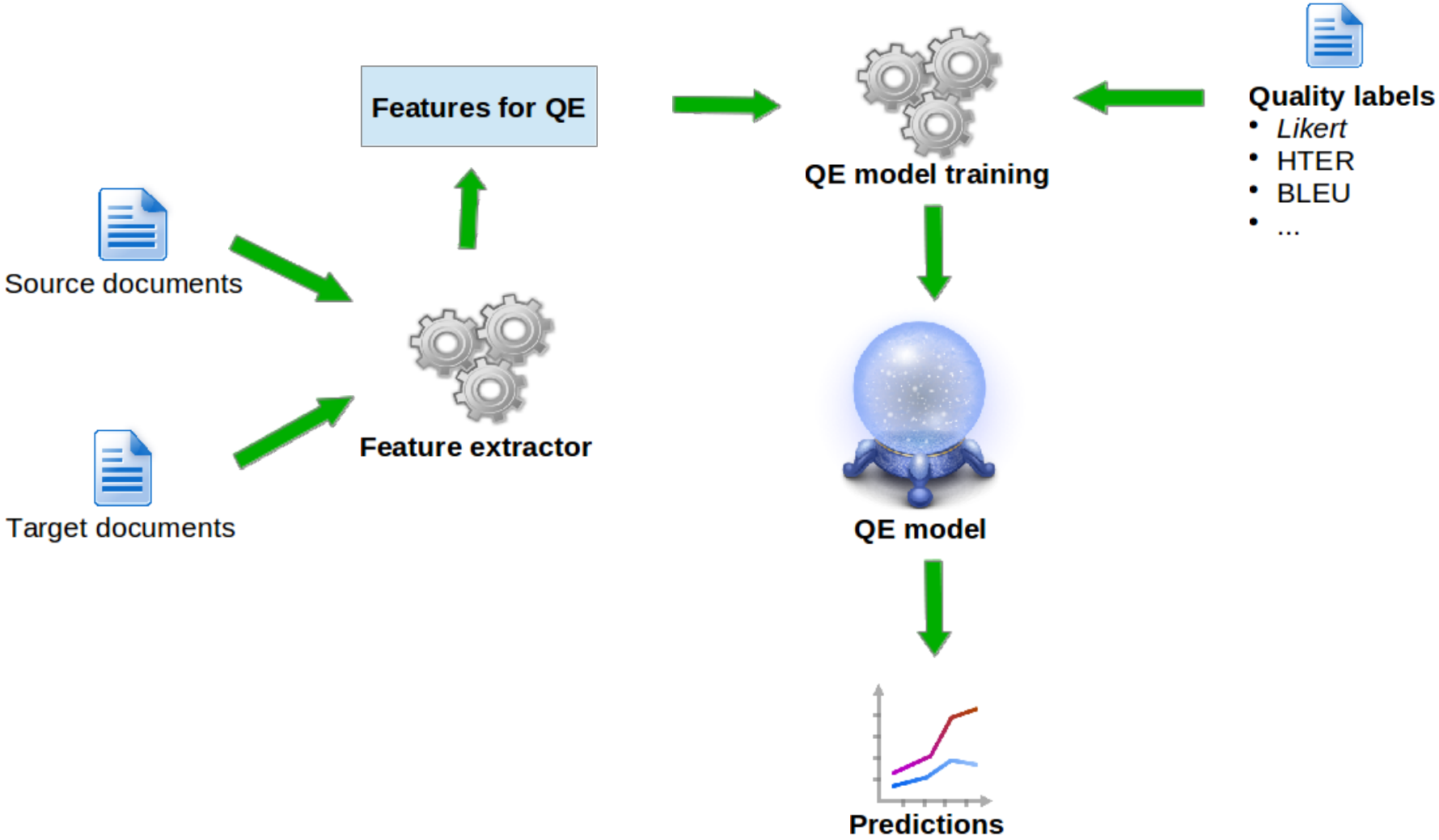
Quality = **Is it worth post-editing it?**

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

Quality = **Which words need fixing?**

Quality = **Which version of the text is more reliable?**

General method

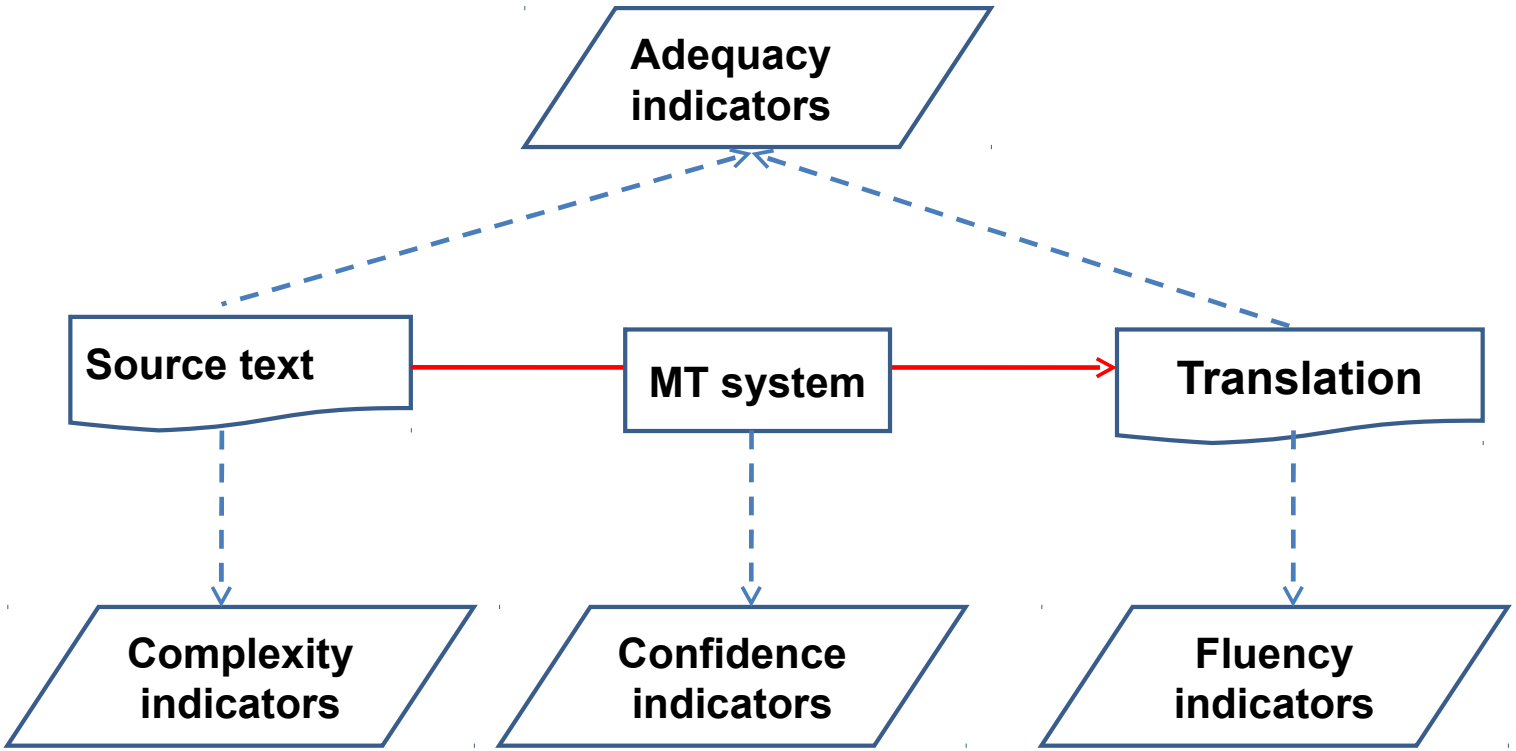


General method

Main components to build a QE system:

1. Definition of quality: **what to predict** and at what **level**
   - Word/phrase
   - Sentence
   - Document
2. (Human) labelled **data** (for quality)
3. **Features**
4. Machine learning **algorithm**

Features


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

Sentence-level QE

- Most popular level
  - MT systems work at sentence-level
  - PE is done at sentence-level
- Easier to get labelled data
- Practical for **post-editing** purposes (edits, time, effort)

Sentence-level QE - Features

MT system-independent features:

- **SF - Source complexity features**:
  - source sentence length
  - source sentence type/token ratio
  - average source word length
  - source sentence 3-gram LM score
  - percentage of source 1 to 3-grams seen in the MT training corpus
  - depth of syntactic tree

- **TF - Target fluency features**:
  - target sentence 3-gram LM score
  - translation sentence length
  - proportion of mismatching opening/closing brackets and quotation marks in translation
  - coherence of the target sentence

Sentence-level QE - Features

- **AF - Adequacy features**:
  - ratio of number of tokens btw source & target and v.v.
  - absolute difference btw no tokens in source & target
  - absolute difference btw no brackets, numbers, punctuation symbols in source & target
  - ratio of no, content-/non-content words btw source & target
  - ratio of nouns/verbs/pronouns/etc btw source & target
  - proportion of dependency relations with constituents aligned btw source & target
  - difference btw depth of the syntactic trees of source & target
  - difference btw no pp/np/vp/adjp/advp/conjp phrase labels in source & target
  - difference btw no 'person'/'location'/'organization' (aligned) entities in source & target
  - proportion of matching base-phrase types at different levels of source & target parse trees

Sentence-level QE - Features

- **Confidence** features:
  - score of the hypothesis (MT global score)
  - size of nbest list
  - using n-best to build LM: sentence n-gram log-probability
  - individual model features (phrase probabilities, etc.)
  - maximum/minimum/average size of the phrases in translation
  - proportion of unknown/untranslated words
  - n-best list density (vocabulary size / average sentence length)
  - edit distance of the current hypothesis to the center hypothesis
  - Search graph info: total hypotheses, % discarded / pruned / recombined search graph nodes
- **Others**:
  - quality prediction for words/phrases in sentence
  - embeddings or other vector representations

Sentence-level QE - Algorithms

- Mostly **regression** algorithms (SVM, GP)
- Binary **classification**: good/bad
- **Kernel** methods perform better
- **Tree kernel** methods for syntactic trees
- **NN** are difficult to train (small datasets)

Sentence-level QE - Predicting HTER @WMT16

Languages, data and MT systems

- 12K/1K/2K train/dev/test English $\rightarrow$ German (**QT21**)
- One SMT system
- IT domain
- Post-edited by professional translators
- Labelling: HTER

Sentence-level QE - Results @WMT16

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

Sentence-level QE - Best results @WMT16

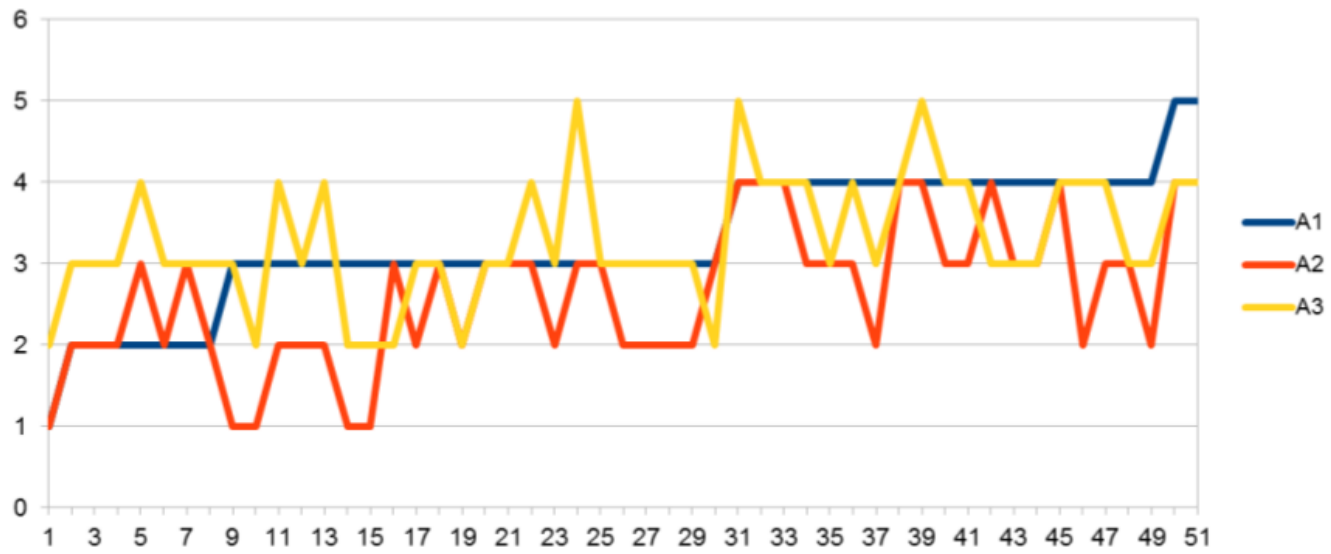
- **YSDA**: features about complexity of source (depth of parse tree, specific constructions), pseudo-reference, back translation, web-scale LM, and word alignments. Trained to predict BLEU scores, followed by a linear SVR to predict HTER from BLEU scores.
- **POSTECH**: RNN with two components: (i) two bidirectional RNNs on the source and target sentences plus (ii) other RNNs for predicting the final quality: (i) is an RNN-based modified NMT model that generates a sequence of vectors about target words' translation quality. (ii) predicts the quality at sentence level. Each component is trained separately: (i) relies on the Europarl parallel corpus, (ii) relies on the QE task data.

Sentence-level QE - Challenges

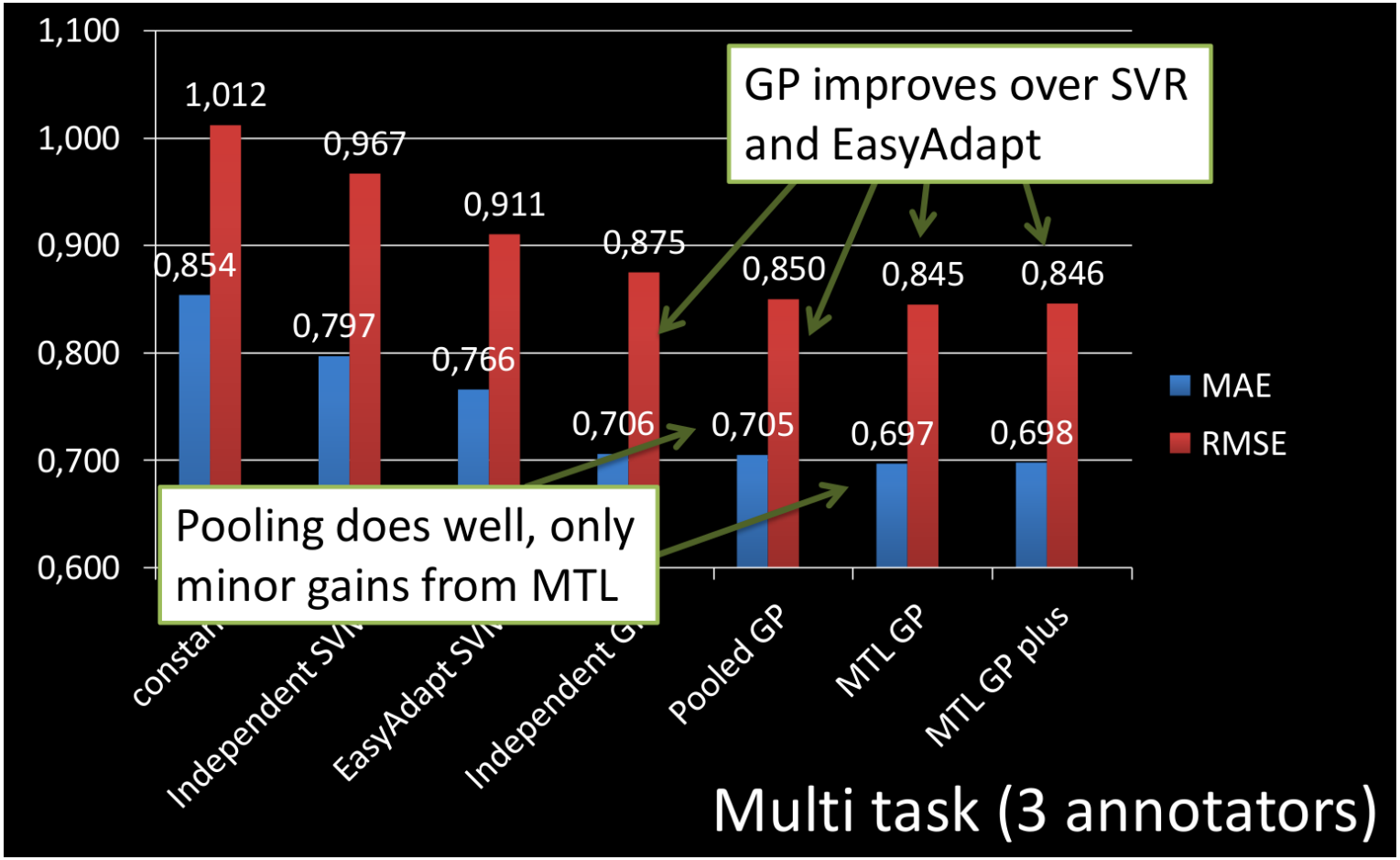
- **Data**: how to obtain objective labels, for different languages and domains, which are comparable across translators?
- How to **adapt** models over time? $\rightarrow$ online learning [C. de Souza et al., 2015]
- How to deal with **biases** from annotators (or domains)? $\rightarrow$ multi-task learning [Cohn and Specia, 2013]

Sentence-level QE - Learning from multiple annotators

- Perception of quality varies
- E.g.: English-Spanish translations labelled for PE effort between 1 (bad) and 5 (perfect)
- 3 annotators: average of 1K scores: **4**; **3.7**; **3.3**

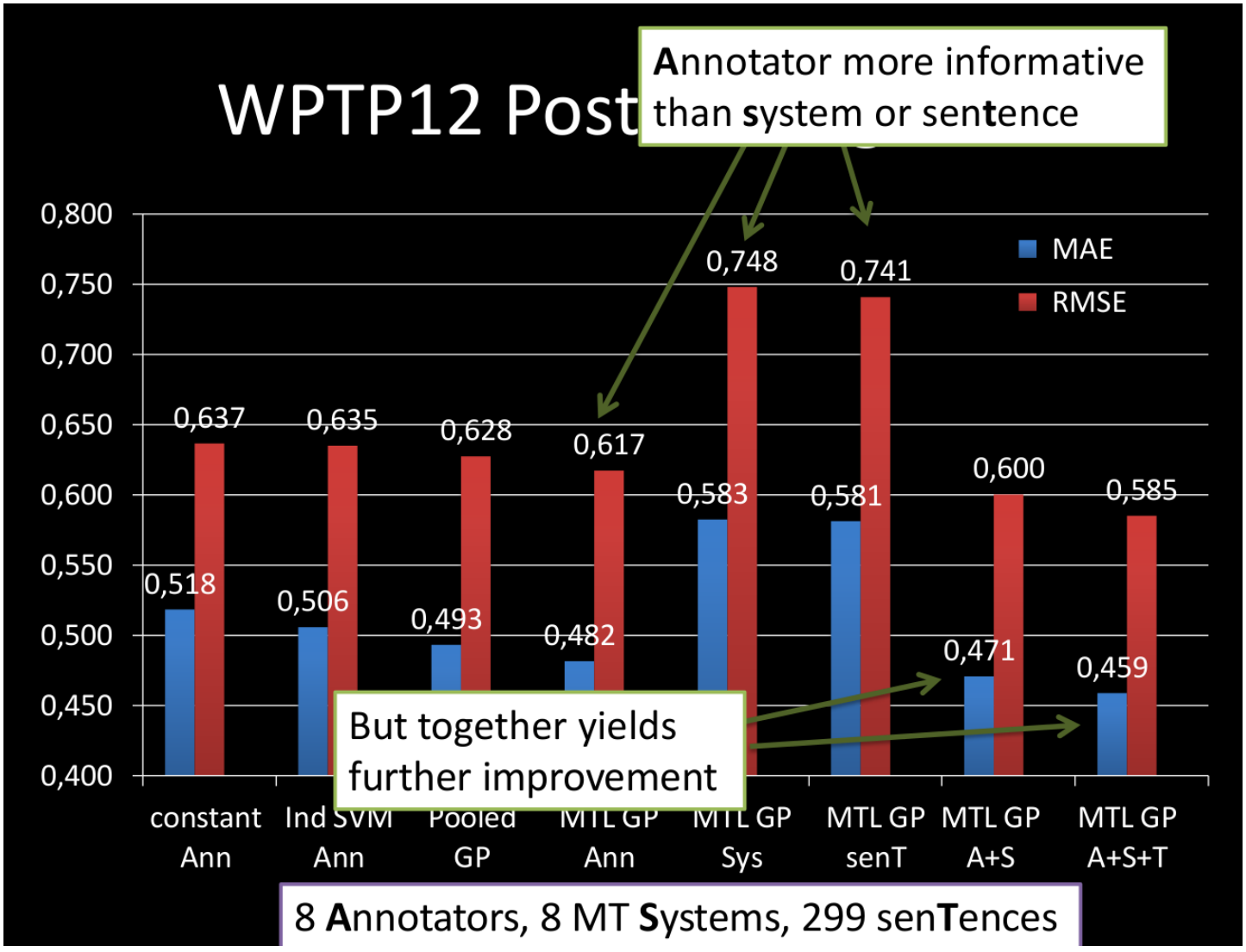


Sentence-level QE - Learning from multiple annotators



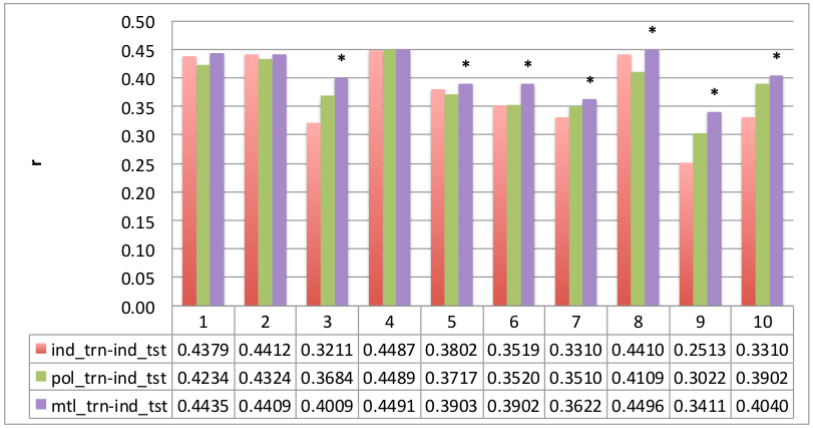
[Cohn and Specia, 2013]

Sentence-level QE - Learning from multiple annotators



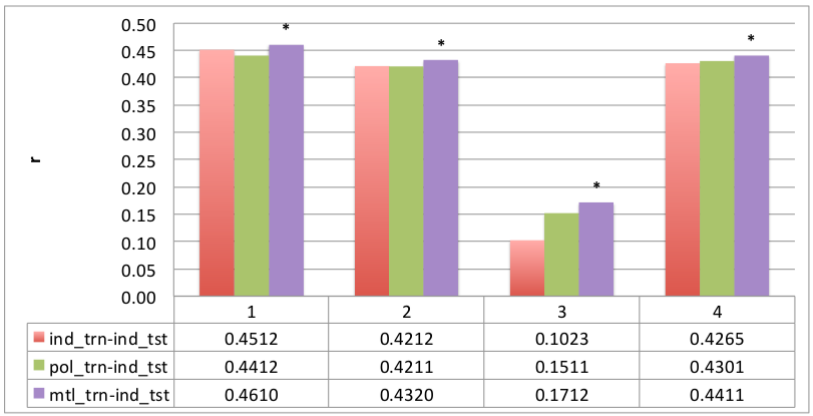
[Cohn and Specia, 2013]

Sentence-level QE - Learning from multiple annotators



	1	2	3	4	5	6	7	8	9	10
ind_trn-ind_tst	0.4379	0.4412	0.3211	0.4487	0.3802	0.3519	0.3310	0.4410	0.2513	0.3310
pol_trn-ind_tst	0.4234	0.4324	0.3684	0.4489	0.3717	0.3520	0.3510	0.4109	0.3022	0.3902
mtl_trn-ind_tst	0.4435	0.4409	0.4009	0.4491	0.3903	0.3902	0.3622	0.4496	0.3411	0.4040

(a) en-es



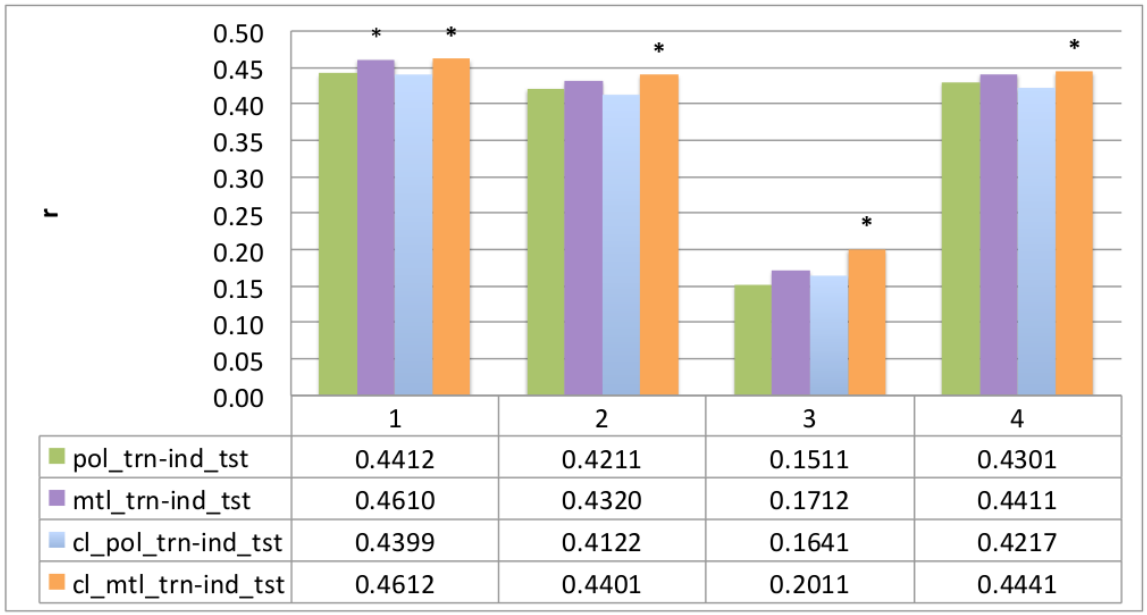
	1	2	3	4
ind_trn-ind_tst	0.4512	0.4212	0.1023	0.4265
pol_trn-ind_tst	0.4412	0.4211	0.1511	0.4301
mtl_trn-ind_tst	0.4610	0.4320	0.1712	0.4411

(b) en-fr

Lang. Pair	ID	Total	Train	Test
en-es	1	28,423	21,317	7,105
	2	12,904	9,678	3,226
	3	3,939	2,954	984
	4	16,518	12,388	4,129
	5	14,187	10,640	3,546
	6	9,395	7,046	2,348
	7	402	301	100
	8	9,294	6,970	2,323
	9	845	633	211
	10	2,756	2,067	689
	All	98,663	73,997	24,665
en-fr	1	65,280	48,960	16,320
	2	6,336	4,752	1,584
	3	769	576	192
	4	5,271	3,953	1,317
	All	77,656	58,241	19,413

[Shah and Specia, 2016]

Sentence-level QE - Learning from multiple annotators



	1	2	3	4
pol_trn-ind_tst	0.4412	0.4211	0.1511	0.4301
mtl_trn-ind_tst	0.4610	0.4320	0.1712	0.4411
cl_pol_trn-ind_tst	0.4399	0.4122	0.1641	0.4217
cl_mtl_trn-ind_tst	0.4612	0.4401	0.2011	0.4441

Predict en-fr using en-fr & en-es [Shah and Specia, 2016]

Word-level QE

Some **applications** require fine-grained information on quality:

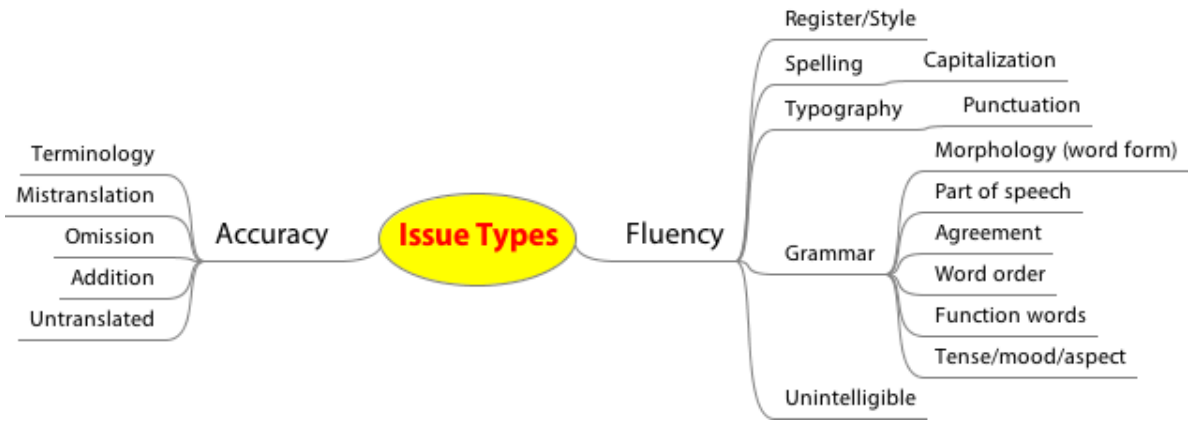
- Highlight words that need fixing
- Inform readers of portions of sentence that are not reliable

Seemingly a **more challenging task**

- A quality label is to be predicted for each target word
- Sparsity is a serious issue
- Skewed distribution towards GOOD
- Errors are interdependent

Word-level QE - Labels

- Predict binary **GOOD/BAD** labels
- Predict general **types of edits**:
  - Shift
  - Replacement
  - Insertion
  - **Deletion** is an issue
- Predict specific errors. E.g. **MQM** in WMT14

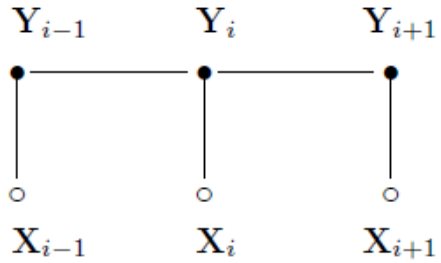


Word-level QE - Features

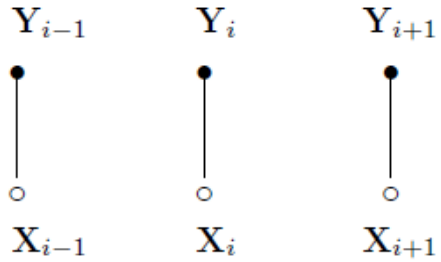
- target token, its left & right token
- source token aligned to target token, its left & right tokens
- boolean dictionary flag: whether target token is a stopword, a punctuation mark, a proper noun, a number
- dangling token flag (null link)
- LM of n-grams with target token t_i: (t_{i-2}, t_{i-1}, t_i), (t_{i-1}, t_i, t_{i+1}), (t_i, t_{i+1}, t_{i+2})
- order of the highest order n-gram which starts/ends with the source/target token
- POS tag of target/source token
- number of senses of target/source token in WordNet
- pseudo-reference flag: 1 if token belongs to pseudo-reference, 0 otherwise

Word-level QE - Algorithms

- **Sequence labelling** algorithms, like **CRF**



- **Classification** algorithms: each word tagged independently



- **NN**:
  - MLP with bilingual word embeddings and standard features
  - RNNs

Word-level QE @WMT16

Languages, data and MT systems

- Same as for T1
- Labelling done with TERCOM:
  - OK = unchanged
  - BAD = insertion, substitution
- Instances: <source word, MT word, OK/BAD label>

	Sentences	Words	% of BAD words
Training	12,000	210,958	21.4
Dev	1,000	19,487	19.54
Test	2,000	34,531	19.31

New **evaluation metric**:

$$F_1\text{-multiplied} = F_1\text{-OK} \times F_1\text{-BAD}$$

Challenge: skewed class distribution

Word-level QE - Results @WMT16

System ID	F_1-**mult** ↑	F_1-BAD	F_1-OK
English-German			
• UNBABEL/ensemble	0.495	0.560	0.885
UNBABEL/linear	0.463	0.529	0.875
UGENT-LT3/SCATE-RF	0.411	0.492	0.836
UGENT-LT3/SCATE-ENS	0.381	0.464	0.821
POSTECH/WORD-RNN-QV3	0.380	0.447	0.850
POSTECH/WORD-RNN-QV2	0.376	0.454	0.828
UAlacant/SBI-Online-baseline	0.367	0.456	0.805
CDACM/RNN	0.353	0.419	0.842
SHEF/SHEF-MIME-1	0.338	0.403	0.839
SHEF/SHEF-MIME-0.3	0.330	0.391	0.845
Baseline CRF	0.324	0.368	0.880
RTM/s5-RTM-GLMd	0.308	0.349	0.882
UAlacant/SBI-Online	0.290	0.406	0.715
RTM/s4-RTM-GLMd	0.273	0.307	0.888
All OK baseline	0.0	0.0	0.893
All BAD baseline	0.0	0.323	0.0

Word-level QE - Results @WMT16

- **Unbabel**: linear sequential model with baseline features + dependency-based features (relations, heads, siblings and grandparents, etc.), and predictions by an ensemble method that uses a stacked architecture which combines three neural systems: one feedforward and two recurrent ones.
- **UGENT**: 41 features + baseline feature set to train binary Random Forest classifiers. Features capture accuracy errors using word and phrase alignment probabilities, fluency errors using language models, and terminology errors using a bilingual terminology list.

Word-level QE - Challenges

- **Data**:
  - Labelling is expensive
  - Labelling from post-editing not reliable → need better alignment methods
  - Data sparsity and skewness are hard to overcome →
    - **Injecting errors** or **filtering positive cases** [Logacheva and Specia, 2015]

- Errors are rarely isolated – how to model **interdependencies**?
  → **Phrase-level QE - WMT16**

Document-level QE

- Prediction of a single label for **entire documents**
- **Assumption**: quality of a document is more than the simple aggregation of its sentence-level quality scores
  - While certain sentences are perfect in isolation, their combination in context may lead to an incoherent document
  - A sentence can be poor in isolation, but good in context as it may benefit from information in surrounding sentences
- **Application**: use as is (no PE) for gisting purposes

Document-level QE - Labels

- Notion of **quality** is very subjective [Scarton and Specia, 2014]
  - Human labels: hard and expensive to obtain, no datasets available
- Most work predicts **METEOR/BLEU** against an independently created reference. Not ideal:
  - Low variance across documents
  - Do not capture discourse issues
- Alternative: task-based labels
  - **2-stage post-editing**
  - Reading comprehension tests

Document-level QE - Features

- average or doc-level counts of **sentence-level features**
- word/lemma/noun repetition in source/target doc and ratio
- number of pronouns in source/target doc
- number of discourse connectives (*expansion*, *temporal*, *contingency*, *comparison* & *non-discourse*)
- number of EDU (elementary discourse units) breaks in source/target doc
- number of RST *nucleus* relations in source/target doc
- number of RST *satellite* relations in source/target doc
- **average quality prediction for sentences in docs**

Algorithms: same as for sentence-level

Document-level QE @WMT16

Languages, data and MT systems

- English $\rightarrow$ Spanish
- Documents by all WMT8-13 translation task MT systems
- 146/62 documents for training/test
- Labelling: 2-stage post-editing method
  1. **PE1**: Sentences are post-edited in arbitrary order (no context)
  2. **PE2**: Post-edited sentences are further edited within document context

Document-level QE @WMT16

Label

- Linear combination of HTER values:

$$w_1 \cdot PE_1 \times MT + w_2 \cdot PE_2 \times PE_1$$

- w_1 and w_2 are learnt empirically to:
  - **Maximise model performance** (MAE/Pearson) and/or
  - **Maximise data variation** (STDEV/AVG)

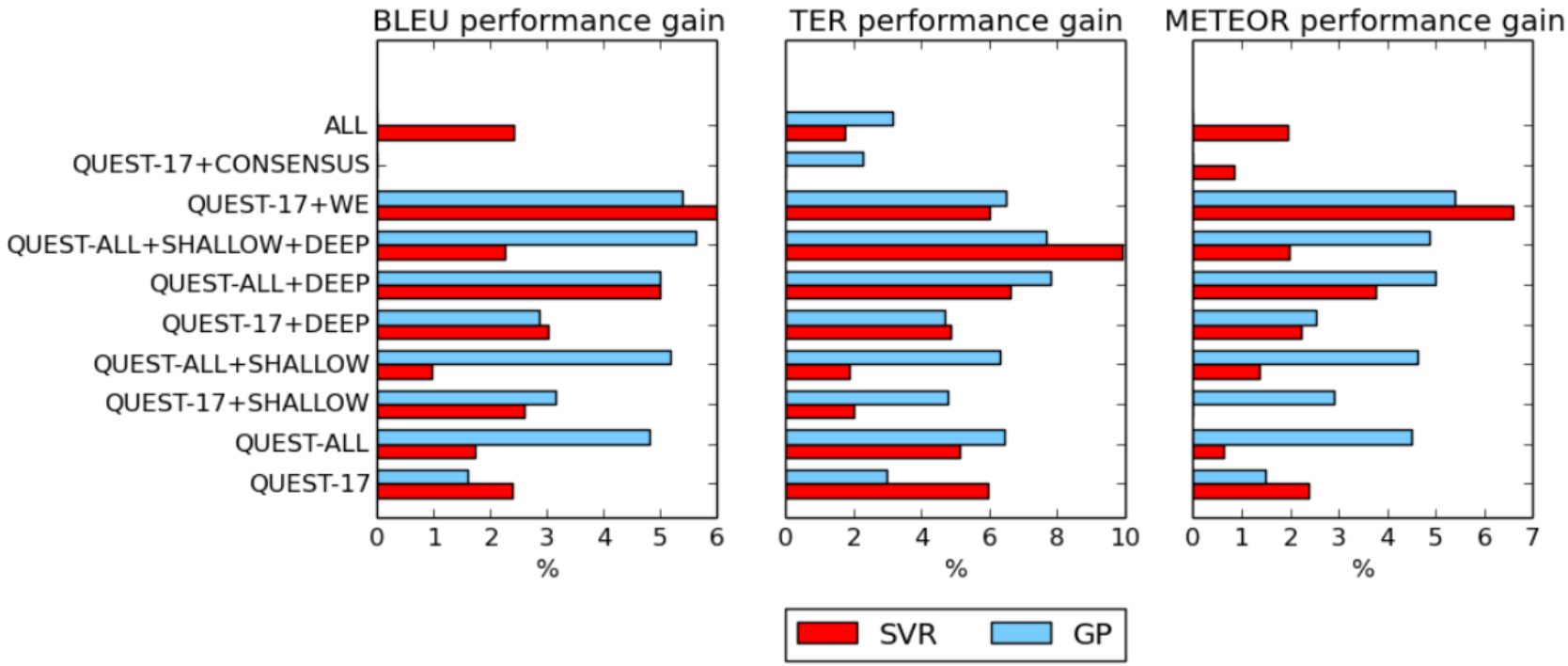
Document-level QE - Results @WMT16

System ID	**Pearson's r**	Spearman's $\rho \uparrow$
English-Spanish		
• **USHEF**/BASE-EMB-GP	0.391	0.393
• RTM/RTM-FS+PLS-TREE	0.356	0.476
RTM/RTM-FS-SVR	0.293	0.360
Baseline SVM	0.286	0.354
USHEF/GRAPH-DISC	0.256	0.285

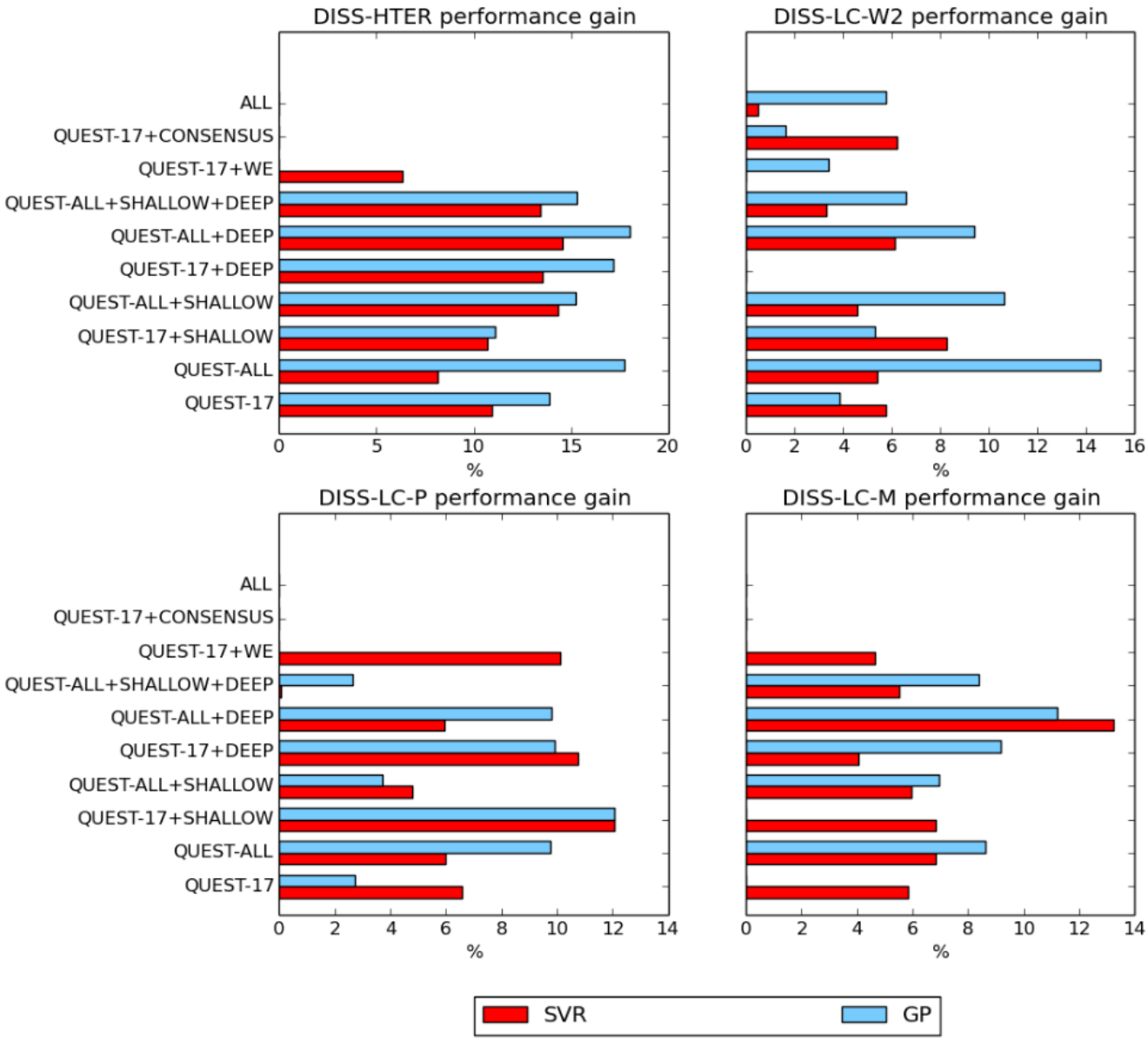
- **USHEF**: 17 baseline features + word embeddings from source documents combined using GP. Document embeddings are the average of the word embeddings in the document. GP model was trained with 2 kernels: one for the 17 baseline features and another for the 500 features from the embeddings.

Document-level QE - New label

MAE gain (%) compared to "mean" baseline:



Document-level QE - New label



Document-level QE - Challenges

- **Quality label** still an open issue
  - should take into account purpose of translation
  - should reliably distinguish different documents
- **Feature engineering**: few tools for discourse processing
  - Topic and structure of document
  - Relationship between its sentences/paragraphs
- **Relevance** information needed: how to factor it in
  - Features: QE for sentence + sentence ranking methods [Turchi et al., 2012]
  - Labels

Participants @WMT16

	WL/PL	SL	DL
Centre for Development of Advanced Computing, India	X		
Pohang University of Science and Technology, Republic of Korea	X	X	
Referential Translation Machines, Turkey	X	X	X
University of Sheffield, UK	X	X	X
University of Sheffield, UK & Lab. d'Informatique de l'Université du Maine, France		X	
University of Alicante, Spain	X		
Nile University, Egypt & Charles University, Czech Republic		X	
Ghent University, Belgium	X	X	
Unbabel, Portugal	X		
Uppsala University, Sweden		X	
Yandex, Russia		X	

Neural Nets for QE

As features:

- NLM for sentence-level
- Word embeddings for word, sentence and doc-level

As learning algorithm:

- MLP proved effective until 2015
- 2016 submissions use RNNs for sentence-level (POSTECH, SimpleNets), word-level (Unbabel), phrase-level (CDAC)

QE in practice

Does QE help?

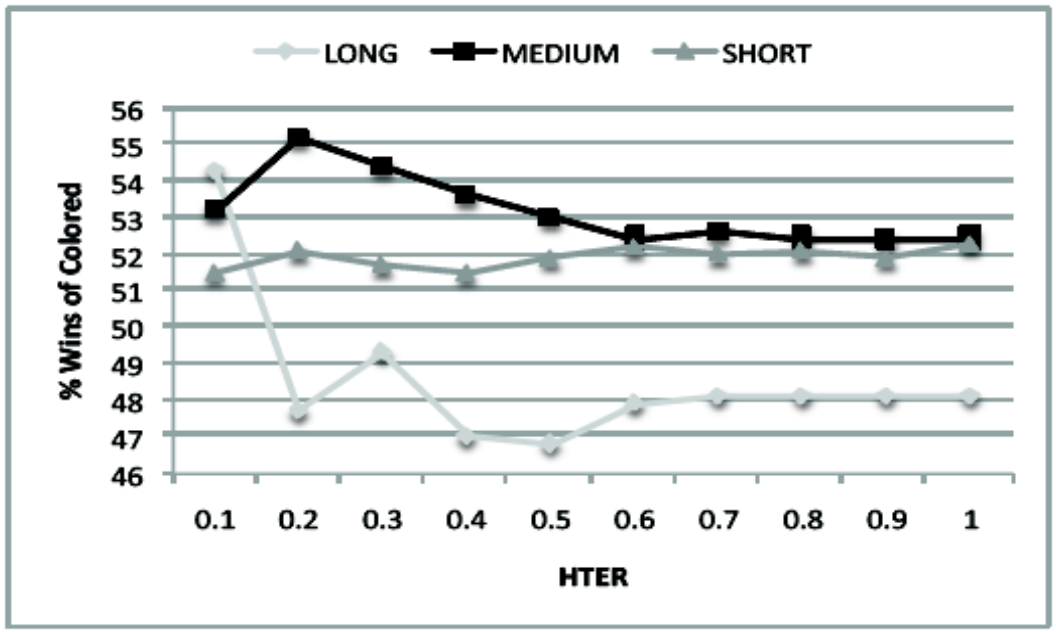
- **Time to post-edit** subset of sentences predicted as "low PE effort" **vs** time to post-edit random subset of sentences [Specia, 2011]

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

QE in practice

- **Productivity increase** [Turchi et al., 2015]
- Comparison btw post-editing with and without QE
- Predictions shown with binary colour codes (**green** vs **red**)

Average PET (sec/word)	colored grey	8.086 9.592	$p = 0.33$
% Wins of colored	51.7		$p = 0.039$



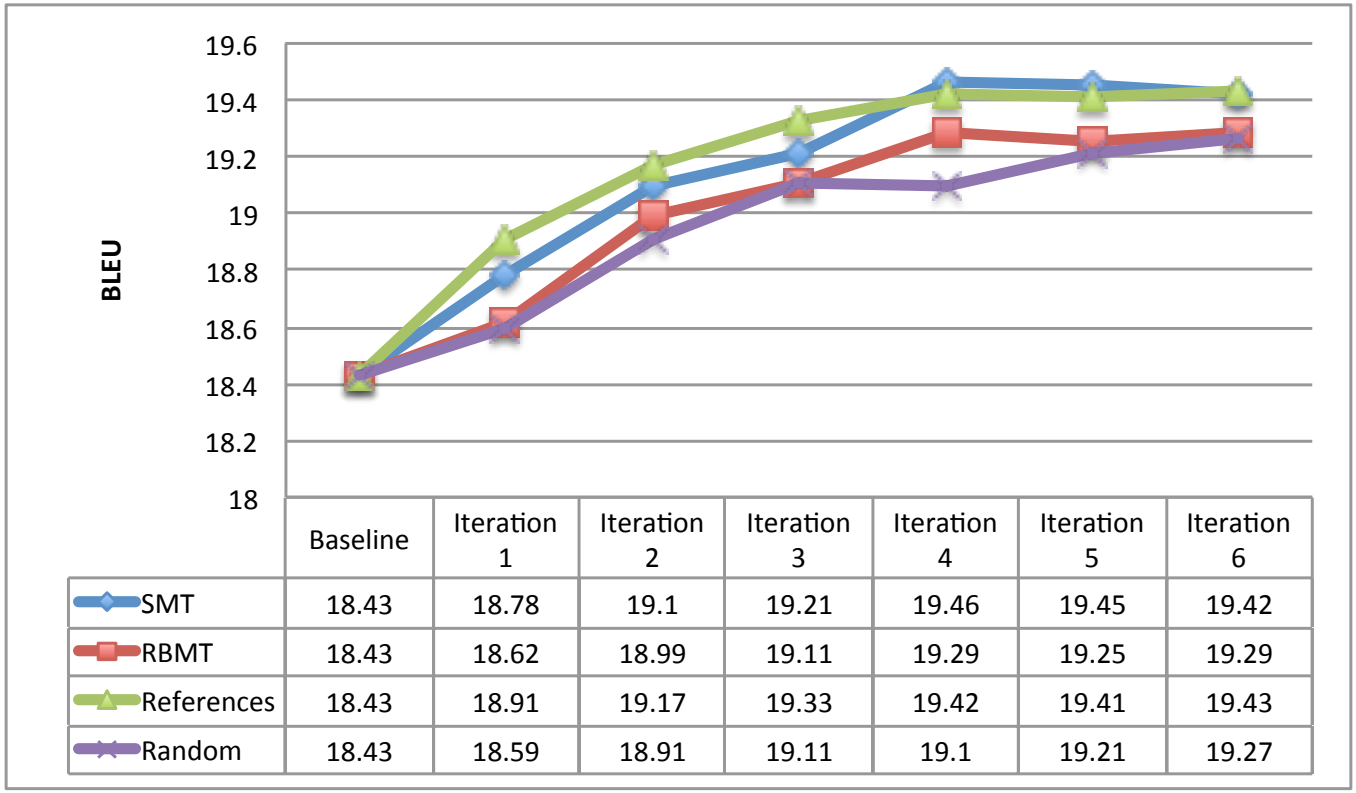
QE in practice

- **MT system selection**: BLEU scores [Specia and Shah, 2016]

	Majority Class	**Best QE-selected**	Best MT system
en-de	16.14	**18.10**	17.04
de-en	25.81	**28.75**	27.96
en-es	30.88	**33.45**	25.89
es-en	30.13	**38.73**	37.83

QE in practice

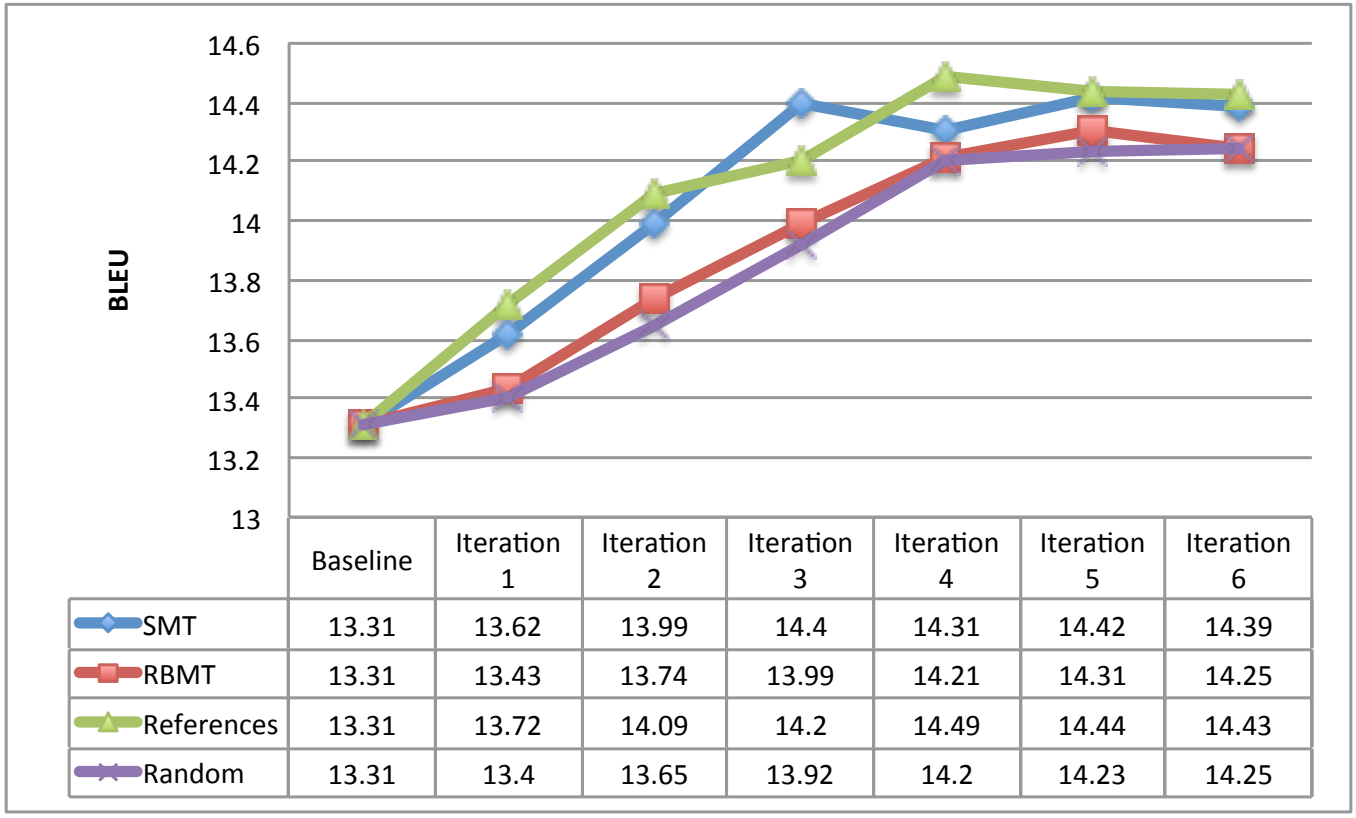
- **SMT self-learning**: **de-en** SMT enhanced with MT data 'best' according to QE [Specia and Shah, 2016]



	Baseline	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5	Iteration 6
SMT	18.43	18.78	19.1	19.21	19.46	19.45	19.42
RBMT	18.43	18.62	18.99	19.11	19.29	19.25	19.29
References	18.43	18.91	19.17	19.33	19.42	19.41	19.43
Random	18.43	18.59	18.91	19.11	19.1	19.21	19.27

QE in practice

- **SMT self-learning**: **en-de** SMT enhanced with MT data 'best' according to QE [Specia and Shah, 2016]



	Baseline	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5	Iteration 6
SMT	13.31	13.62	13.99	14.4	14.31	14.42	14.39
RBMT	13.31	13.43	13.74	13.99	14.21	14.31	14.25
References	13.31	13.72	14.09	14.2	14.49	14.44	14.43
Random	13.31	13.4	13.65	13.92	14.2	14.23	14.25

Quality Estimation for Language Output Applications

Carolina Scarton, Gustavo Paetzold and Lucia Specia

University of Sheffield, UK



COLING, Osaka, 11 Dec 2016

References I

C. de Souza, J. G., Negri, M., Ricci, E., and Turchi, M. (2015).

Online multitask learning for machine translation quality estimation.

In 53rd Annual Meeting of the Association for Computational Linguistics, pages 219–228, Beijing, China.

Cohn, T. and Specia, L. (2013).

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, pages 32–42, Sofia, Bulgaria.

Logacheva, V. and Specia, L. (2015).

The role of artificially generated negative data for quality estimation of machine translation.

In 18th Annual Conference of the European Association for Machine Translation, EAMT, Antalya, Turkey.

References II

Louis, A. and Nenkova, A. (2013).

Automatically assessing machine summary content without a gold standard.

Computational Linguistics, 39(2):267–300.

Ravi, S., Knight, K., and Soricut, R. (2008).

Automatic prediction of parser accuracy.

In Conference on Empirical Methods in Natural Language Processing, pages 887–896, Honolulu, Hawaii.

Scarton, C. and Specia, L. (2014).

Document-level translation quality estimation: exploring discourse and pseudo-references.

In 17th Annual Conference of the European Association for Machine Translation, EAMT, pages 101–108, Dubrovnik, Croatia.

References III

Shah, K. and Specia, L. (2016).

Large-scale multitask learning for machine translation quality estimation.

In Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 558–567, San Diego, California.

Singh, A. and Jin, W. (2016).

Ranking Summaries for Informativeness and Coherence without Reference Summaries.

In The Twenty-Ninth International Florida Artificial Intelligence Research Society Conference, pages 104–109, Key Largo, Florida.

Specia, L. (2011).

Exploiting objective annotations for measuring translation post-editing effort.

In 15th Conference of the European Association for Machine Translation, pages 73–80, Leuven.

References IV

Specia, L. and Shah, K. (2016).

Machine Translation Quality Estimation: Applications and Future Perspectives, page to appear.

Springer.

Turchi, M., Negri, M., and Federico, M. (2015).

Mt quality estimation for computer-assisted translation: Does it really help?

In 53rd Annual Meeting of the Association for Computational Linguistics, pages 530–535, Beijing, China.

Turchi, M., Specia, L., and Steinberger, J. (2012).

Relevance ranking for translated texts.

In 16th Annual Conference of the European Association for Machine Translation, EAMT, pages 153–160, Trento, Italy.