5th Quality Estimation Shared Task

WMT16

Lucia Specia, Varvara Logacheva and Carolina Scarton

University of Sheffield

Berlin, 12 August 2016

Outline

Goals

QE metrics predict the quality of a translated text **without a reference translation**

Goals in 2016:

- Advance work on sentence and word-level QE
  - High quality datasets, professionally post-edited
- Introduce a phrase-level task
- Introduce a document-level task

Tasks

- T1: Predicting **sentence-level** post-editing (PE) distance
- T2: Predicting **word** and **phrase-level** OK/BAD labels
- T3: Predicting **document-level** 2-stage PE distance

Participants

ID	Team
CDACM	Centre for Development of Advanced Computing, India
POSTECH	Pohang University of Science and Technology, Republic of Korea
RTM	Referential Translation Machines, Turkey
SHEF	University of Sheffield, UK
SHEF-LIUM	University of Sheffield, UK and Laboratoire d'Informatique de l'Université du Maine, France
SHEF-MIME	University of Sheffield, UK
UAlacant	University of Alicante, Spain
UFAL	Nile University, Egypt & Charles University, Czech Republic
UGENT	Ghent University, Belgium
UNBABEL	Unbabel, Portugal
USFD	University of Sheffield, UK
USHEF	University of Sheffield, UK
UU	Uppsala University, Sweden
YSDA	Yandex, Russia

14 teams, **39 systems**: up to 2 per team, per subtask

Outline

Predicting sentence-level HTER

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
- Instances: <SRC, MT, PE, HTER>

Predicting sentence-level HTER

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	–

• = winning submissions - top-scoring and those which are not significantly worse.
Gray area = systems that are not significantly different from the baseline.

Predicting sentence-level HTER: 2016 vs 2015

Different language pair, different domain, different MT system:

System ID (**2015**)	**Pearson's r ↑**
English-Spanish	
• LORIA/17+LSI+MT+FILTRE	0.39
• LORIA/17+LSI+MT	0.39
• RTM-DCU/RTM-FS+PLS-SVR	0.38
RTM-DCU/RTM-FS-SVR	0.38
UGENT-LT3/SCATE-SVM	0.37
UGENT-LT3/SCATE-SVM-single	0.32
SHEF/SVM	0.29
SHEF/GP	0.19
Baseline SVM	0.14

Outline

Predicting word-level quality

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

Challenge: skewed class distribution

Predicting word-level quality

- Mostly interested in finding errors
- Precision/recall preferences depend on application
- Rare classes should not dominate

New **evaluation metric**:

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

Baseline:

- CRF classifier with 22 features

Predicting word-level quality

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

Predicting word-level quality: 2016 vs 2015

System ID (2015)	F_1-mult	F_1-**BAD**	F_1-OK
English-Spanish			
• UAlacant/OnLine-SBI-Baseline	0.336	0.431	0.781
• HDCL/QUETCHPLUS	0.342	0.431	0.794
UAlacant/OnLine-SBI	0.316	0.415	0.761
SAU/KERC-CRF	0.338	0.391	0.864
SAU/KERC-SLG-CRF	0.336	0.389	0.864
SHEF2/W2V-BI-2000	0.275	0.384	0.716
SHEF2/W2V-BI-2000-SIM	0.275	0.384	0.715
SHEF1/QuEst++-AROW	0.259	0.384	0.676
UGENT/SCATE-HYBRID	0.305	0.367	0.830
DCU-SHEFF/BASE-NGRAM-2000	0.273	0.366	0.745
HDCL/QUETCH	0.298	0.353	0.846
DCU-SHEFF/BASE-NGRAM-5000	0.292	0.345	0.845
SHEF1/QuEst++-PA	0.836	0.343	0.244
All BAD baseline	0.00	0.318	0.00
UGENT/SCATE-MBL	0.258	0.306	0.843
RTM-DCU/s5-RTM-GLMd	0.211	0.239	0.881
RTM-DCU/s4-RTM-GLMd	0.200	0.227	0.883
Baseline CRF	0.147	0.168	0.889
All OK baseline	0.00	0.00	0.896

Predicting word-level quality: 2016 vs 2015

- Improved baseline
- New metric: trivial baselines at the bottom
- Better systems: all submissions outperform **all BAD baseline**, even in terms of F_1-BAD

Outline

Predicting phrase-level quality

Languages, data and MT systems

- Same as for T1
- Labelling: TERCOM + phrase segmentation

OK	OK	OK	OK	BAD	BAD	BAD	OK
Beim	Schließen ‖	eines	Dokuments ‖	werden ‖	die	Historie	.
OK		**OK**		**BAD**	**BAD**		

- Instances: <source phrase, MT phrase, OK/BAD label>

	Sentences	Phrases	% of BAD phrases
Training	12,000	109,921	29.84
Dev	1,000	9,024	30.21
Test	2,000	16,450	29.53

Predicting phrase-level quality

Languages, data and MT systems

- Same as for T1
- Labelling: TERCOM + phrase segmentation

OK	OK	OK	OK	BAD	BAD	BAD	OK
Beim	Schließen ‖	eines	Dokuments ‖	werden ‖	die	Historie	.
OK		**OK**		**BAD**	**BAD**		

- Instances: <source phrase, MT phrase, OK/BAD label>

	Sentences	Phrases	% of BAD phrases
Training	12,000	109,921	29.84
Dev	1,000	9,024	30.21
Test	2,000	16,450	29.53

Predicting phrase-level quality

Languages, data and MT systems

- Same as for T1
- Labelling: TERCOM + phrase segmentation

OK OK OK OK BAD BAD BAD OK

Beim Schließen || eines Dokuments || werden || die Historie .

OK **OK** **BAD** **BAD**

- Instances: <source phrase, MT phrase, OK/BAD label>

	Sentences	Phrases	% of BAD phrases
Training	12,000	109,921	29.84
Dev	1,000	9,024	30.21
Test	2,000	16,450	29.53

Predicting phrase-level quality

System ID	F_1-**mult** ↑	F_1-BAD	F_1-OK
English-German			
• CDACM/RNN	0.380	0.503	0.755
• POSTECH/PHR-RNN-QV3	0.378	0.495	0.764
• POSTECH/PHR-RNN-QV2	0.369	0.478	0.772
• USFD2/W&SLP4PT	0.368	0.486	0.757
• USFD2/CONTEXT	0.365	0.470	0.777
RTM/s5_RTM-GLMd	0.327	0.408	0.802
Baseline CRF	0.321	0.401	0.800
RTM/s4_RTM-GLMd	0.307	0.377	0.814
Ualacant/SBI-Online-baseline	0.259	0.493	0.526
UAlacant/SBI-Online	0.098	0.459	0.213
All BAD baseline	0.0	0.457	0.0
All OK baseline	0.0	0.0	0.825

Outline

Predicting 2-stage post-editing distance

Languages, data and MT systems

- English → Spanish
- **Whole** documents by all news translation task MT systems (WMT08-13)
- 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

Predicting 2-stage post-editing distance

New 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 $\rightarrow$ **minimise error** (MAE) and **maximise variation** (STDEV/AVG)

	$PE_1 \times MT$	$PE_2 \times PE_1$	**NEW LABEL**
AVG	0.346	0.042	0.895
STDEV	0.108	0.034	0.457
Ratio	0.312	0.810	0.511

Predicting 2-stage post-editing distance

System ID	**Pearson's r**	Spearman's ρ ↑
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

Outline

Discussion

- Steady participation
- Absolute improvements wrt 2015 may be due to **more consistent**, **more repetitive** data
- Best **sentence** and **word-level** systems **by companies**
- **Phrase-level**: more work needed on evaluation
- **Document-level**: few participants, more challenging task?

Discussion

- Steady participation
- Absolute improvements wrt 2015 may be due to **more consistent**, **more repetitive** data
- Best **sentence** and **word-level** systems **by companies**
- **Phrase-level**: more work needed on evaluation
- **Document-level**: few participants, more challenging task?

- Systems doing well in general:

Sentence level	$11 >$	Baseline	> 2
Word level	$10 >$	Baseline	≥ 3
Phrase level	$5 >$	Baseline	≥ 4
Document level	$2 >$	Baseline	≥ 2

Next round

- Larger datasets (**QT21**): 45K segments
- EN-DE/DE-EN and potentially other language pairs
- Continue with traditional variants
  - More on **phrase level**
  - Not sure about **document level**
- Word/phrase-level: beyond OK/BAD

Next round

- Larger datasets (**QT21**): 45K segments
- EN-DE/DE-EN and potentially other language pairs
- Continue with traditional variants
  - More on **phrase level**
  - Not sure about **document level**
- Word/phrase-level: beyond OK/BAD

QuEst: `www.dcs.shef.ac.uk/~quest`

Next round

- Larger datasets (**QT21**): 45K segments
- EN-DE/DE-EN and potentially other language pairs
- Continue with traditional variants
  - More on **phrase level**
  - Not sure about **document level**
- Word/phrase-level: beyond OK/BAD

QuEst: `www.dcs.shef.ac.uk/~quest`

Tutorial on Quality Estimation at **COLING**