4th Quality Estimation Shared Task

WMT15

Lucia Specia[†], Chris Hokamp[§], Varvara Logacheva[†] and Carolina Scarton[†]

[†]University of Sheffield
[§]Dublin City University

Lisbon, 18 September 2015

Outline

Goals in 2015

- Advance work on sentence and word-level QE
  - Larger datasets, but crowdsourced post-editions
  - Same data as for APE task
- Investigate effectiveness of quality labels, features and learning methods for document-level QE
  - Paragraphs as "documents"

Tasks

- T1: Predicting sentence-level edit distance (HTER)
- T2: Predicting word-level OK/BAD labels
- T3: Predicting paragraph-level Meteor

Participants

ID	Team
DCU-SHEFF	Dublin City University, Ireland and University of Sheffield, UK
HDCL	Heidelberg University, Germany
LORIA	Lorraine Laboratory of Research in Computer Science and its Applications, France
RTM-DCU	Dublin City University, Ireland
SAU-KERC	Shenyang Aerospace University, China
SHEFF-NN	University of Sheffield Team 1, UK
UAlacant	Alicant University, Spain
UGENT	Ghent University, Belgium
USAAR-USHEF	University of Sheffield, UK and Saarland University, Germany
USHEF	University of Sheffield, UK
HIDDEN	Undisclosed

10 teams, **34 systems**: up to 2 per team, per subtask

Outline

Predicting sentence-level HTER

Languages and MT systems

- English $\rightarrow$ Spanish
- One MT system
- News
- Training: $12,271$ <source, MT, PE, HTER>
- Test: $1,817$ <source, MT>

Predicting sentence-level HTER

System ID	**MAE ↓**
English-Spanish	
• RTM-DCU/RTM-FS+PLS-SVR	13.25
• LORIA/17+LSI+MT+FILTRE	13.34
• RTM-DCU/RTM-FS-SVR	13.35
• LORIA/17+LSI+MT	13.42
• UGENT-LT3/SCATE-SVM	13.71
UGENT-LT3/SCATE-SVM-single	13.76
SHEF/SVM	13.83
Baseline SVM	14.82
SHEF/GP	15.16

• = 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

Did we do better than last year?

System ID	**MAE** ↓
English-Spanish	
• FBK-UPV-UEDIN/WP	12.89
• RTM-DCU/RTM-SVR	13.40
• USHEFF	13.61
RTM-DCU/RTM-TREE	14.03
DFKI/SVR	14.32
FBK-UPV-UEDIN/NOWP	14.38
SHEFF-lite/sparse	15.04
MULTILIZER	15.04
Baseline SVM	15.23
DFKI/SVRxdata	16.01
SHEFF-lite	18.15

Predicting sentence-level HTER

Pearson correlation (Graham, 2015) = **DeltaAvg's ranking**

System ID	**Pearson's r ↑**
• 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 and MT systems - same as for T1

- English → Spanish, one MT system, News
- Labelling done with TERCOM:
  - OK = unchanged
  - BAD = insertion, substitution
- Data: <source word, MT word, OK/BAD label>

	Sentences	Words	% of BAD words
Training	12,271	280,755	19.16
Test	1,817	40,899	18.87

Challenge: skewed class distribution

Predicting word-level quality

Evaluation metric: average $F1$ of "BAD" class

- Mostly interested in finding errors

Baseline introduced

- CRF classifier with 25 features

Predicting word-level quality

System ID	weighted F_1 All ↑	F_1 **BAD** ↑	F_1 OK ↑
English-Spanish			
• UAlacant/OnLine-SBI-Baseline	71.47	43.12	78.07
• HDCL/QUETCHPLUS	72.56	43.05	79.42
UAlacant/OnLine-SBI	69.54	41.51	76.06
SAU/KERC-CRF	77.44	39.11	86.36
SAU/KERC-SLG-CRF	77.4	38.91	86.35
SHEF2/W2V-BI-2000	65.37	38.43	71.63
SHEF2/W2V-BI-2000-SIM	65.27	38.40	71.52
SHEF1/QuEst++-AROW	62.07	38.36	67.58
UGENT/SCATE-HYBRID	74.28	36.72	83.02
DCU-SHEFF/BASE-NGRAM-2000	67.33	36.60	74.49
HDCL/QUETCH	75.26	35.27	84.56
DCU-SHEFF/BASE-NGRAM-5000	75.09	34.53	84.53
SHEF1/QuEst++-PA	26.25	34.30	24.38
Baseline (always BAD)	0.599	31.76	0.00
UGENT/SCATE-MBL	74.17	30.56	84.32
RTM-DCU/s5-RTM-GLMd	76.00	23.91	88.12
RTM-DCU/s4-RTM-GLMd	75.88	22.69	88.26
Baseline CRF	75.31	16.78	88.93
Baseline (always OK)	72.67	0.00	89.58

Predicting word-level quality

How does it compare to last year?

System ID	weighted F_1 All ↑	F_1 **BAD** ↑
Baseline (always BAD)	18.71	**52.53**
• FBK-UPV-UEDIN/RNN	62.00	**48.73**
LIMSI/RF	60.55	47.32
LIG/FS	63.55	44.47
LIG/BL ALL	63.77	44.11
FBK-UPV-UEDIN/CRF	62.17	42.63
RTM-DCU/RTM-GLM	60.68	35.08
RTM-DCU/RTM-GLMd	60.24	32.89
Baseline (always OK)	50.43	0.00

Outline

Predicting paragraph-level Meteor

MT1: According to the specifications this headset supports Bluetooth 1.2. With fashion and Ericsson W600i Sony Walkman, when I was called up when people were tied to **them** (their) mobile phone, who could hear me. I tried every possible configuration, read the instructional leaflets for each device, but the thing does not do anything when connected.

MT2: According to the specifications, this headset, as well as Bluetooth 1.2. I could not make any sound to come out when connected to my Sony Ericsson w600i in mobile phones and Walkman mode, and when I call them, people could not listen me. I have tried all the settings, can read the education booklet for each device, and things will not yet in connection.

Which MT is worse?

Predicting paragraph-level Meteor

Languages and MT systems

- English → German, German → English
- Paragraphs from all WMT13 translation task MT systems
- 800 for training; 415 for test
- Average Meteor scores in data:

	EN-DE		DE-EN	
	AVG	STDEV	AVG	STDEV
Meteor (↑)	0.35	0.14	0.26	0.09

Predicting paragraph-level Meteor

System ID	**MAE** ↓
English-German	
• RTM-DCU/RTM-FS-SVR	7.28
• RTM-DCU/RTM-SVR	7.5
USAAR-USHEF/BFF	9.37
USHEF/QUEST-DISC-REP	9.55
Baseline SVM	10.05
German-English	
• RTM-DCU/RTM-FS-SVR	4.94
RTM-DCU/RTM-FS+PLS-SVR	5.78
USHEF/QUEST-DISC-BO	6.54
USAAR-USHEF/BFF	6.56
Baseline SVM	7.35

Predicting paragraph-level Meteor

Pearson correlation (Graham, 2015) = **DeltaAvg's ranking**

System ID	**Pearson's r ↑**
English-German	
• RTM-DCU/RTM-SVR	0.59
RTM-DCU/RTM-FS-SVR	0.53
USHEF/QUEST-DISC-REP	0.30
USAAR-USHEF/BFF	0.29
Baseline SVM	0.12
German-English	
• RTM-DCU/RTM-FS-SVR	0.52
RTM-DCU/RTM-FS+PLS-SVR	0.39
USHEF/QUEST-DISC-BO	0.10
USAAR-USHEF/BFF	0.08
Baseline SVM	0.06

Outline

Advances in sentence- and word-level QE

- Better sentence and word-level results than WMT14
  - Resources for baseline features less useful this year (?)
- Improvement may have been due to larger training sets, despite potential drop in quality
- For word level, proportion of BAD words was too small:
  - 15% sentences with 0 BAD words
  - 35% sentences with fewer than 15% BAD words
  - Training data manipulation strategies led to improved results: filtering, insertion of additional BAD words

Labels, features & learning for document-level QE

Is it different from sentence-level QE?

- Similar framework: same algorithms, mostly same features
- Few discourse-aware features showed only marginal improvements wrt baseline
  - Very short paragraphs
- "Mean" of training score is a good predictor
  - Same as baseline system
- Adequate quality label for entire document still open issue

Next round

- Sentence and word-level:
  - Large datasets collected as part of **QT21**
  - EN-DE as starting point
  - Professional post-editing and error (MQM) annotation
- **Document level**: new labelling scheme by humans
- Introduction of a **phrase-level** prediction task

Next round

- Sentence and word-level:
  - Large datasets collected as part of **QT21**
  - EN-DE as starting point
  - Professional post-editing and error (MQM) annotation
- **Document level**: new labelling scheme by humans
- Introduction of a **phrase-level** prediction task

Tool used for all tasks: QuEst++ (ACL-demo, 2015), `https://github.com/ghpaetzold/questplusplus`

4th Quality Estimation Shared Task

WMT15

Lucia Specia[†], Chris Hokamp[§], Varvara Logacheva[†] and Carolina Scarton[†]

[†]University of Sheffield
[§]Dublin City University

Lisbon, 18 September 2015

Predicting word-level quality

New metric: **Sequence Correlation**

Reference:	OK	BAD	OK	OK	OK
Hypothesis:	OK	OK	OK	OK	OK

Precision $= 4/5 = 0.8$

Reference:	"OK"	"BAD"	"OK	OK	OK"
Hypothesis:	"OK	OK	OK	OK	OK"

Use each overlapping sequence once: Precision $= 3/5 = 0.6$
and λ_t weigh each tag t inversely proportional to the number of those tags in the reference: $\lambda_{GOOD} = 5/4$ and $\lambda_{BAD} = 5/1$

Predicting word-level quality

System ID	**Sequence Correlation ↑**
English-Spanish	
• SAU/KERC-CRF	34.22
• SAU/KERC-SLG-CRF	34.09
• UAlacant/OnLine-SBI-Baseline	33.84
UAlacant/OnLine-SBI	32.81
HDCL/QUETCH	32.13
HDCL/QUETCHPLUS	31.38
DCU-SHEFF/BASE-NGRAM-5000	31.23
UGENT/SCATE-HYBRID	30.15
DCU-SHEFF/BASE-NGRAM-2000	29.94
UGENT/SCATE-MBL	28.43
SHEF2/W2V-BI-2000	27.65
SHEF2/W2V-BI-2000-SIM	27.61
SHEF1/QuEst++-AROW	27.36
RTM-DCU/s5-RTM-GLMd	25.92
SHEF1/QuEst++-PA	25.49
RTM-DCU/s4-RTM-GLMd	24.95
Baseline CRF	0.2044