Recurrent Neural Network based Translation Quality Estimation

Hyun Kim
Creative IT Engineering,
Pohang University of Science and
Technology (POSTECH),
Pohang, Republic of Korea
hkim.postech@gmail.com

Jong-Hyeok Lee
Computer Science and Engineering,
Pohang University of Science and
Technology (POSTECH),
Pohang, Republic of Korea
jhlee@postech.ac.kr

Abstract

This paper describes the recurrent neural network based model for translation quality estimation. Recurrent neural network based quality estimation model consists of two parts. The first part using two bidirectional recurrent neural networks generates the quality information about whether each word in translation is properly translated. The second part using another recurrent neural network predicts the final quality of translation. We apply this model to sentence, word and phrase level of WMT16 Quality Estimation Shared Task. Our results achieve the excellent performance especially in sentence and phrase-level QE.

1 Introduction

We introduce the recurrent neural network based quality estimation (QE) model for predicting the sentence, word and phrase-level translation qualities, without relying on manual efforts to find QE related features.

Existing QE researches have been usually focused on finding desirable QE related features to use machine learning algorithms. Recently, however, there have been efforts to apply neural networks to QE and these neural approaches have shown potential for QE. Shah et al. (2015) use continuous space language model features for sentence-level QE and word embedding features for word-level QE, in combination with other features produced by QuEst++ (Specia et al., 2015). Kreutzer et al. (2015) apply neural networks using pre-trained alignments and word lookup-table to word-level QE, which achieve the excellent performance by using the combination of baseline features at word level. However, these are not 'pure' neural approaches for QE.

Kim and Lee (2016) apply neural machine translation (NMT) models, based on recurrent neural network, to sentence-level QE. This is the first try of using NMT models for the translation quality estimation. This recurrent neural network based quality estimation model is a pure neural approach for QE and achieves a competitive performance in sentence-level QE (English-Spanish).

In this paper, we extend the recurrent neural network based quality estimation model to word and phrase level. Also, we apply this model to sentence, word and phrase-level QE shared task (English-German) of WMT16.

2 Recurrent Neural Network based Quality Estimation Model

Recurrent neural network (RNN) based quality estimation model (Kim and Lee, 2016) consists of two parts: two bidirectional RNNs on the source and target sentences in the first part and another RNN for predicting the quality in the second part.

In the first part (Figure 1), modified RNN-based NMT model generates *quality vectors*, which indicate a sequence of vectors about target words' translation qualities. Each quality vector for each target word has, as not a number unit but a vector unit, the quality information about whether each target word is properly translated from source sentence. Each quality vector is generated by decomposing the probability of each target word from the modified NMT model.[1] Kim and Lee (2016) modify the NMT model to 1) use source and target

[1] Existing NMT models (Cho et al., 2014; Bahdanau et al., 2015) use RNNs on source and target sentences to predict the probability of target word.

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 787–792,
Berlin, Germany, August 11-12, 2016. ©2016 Association for Computational Linguistics

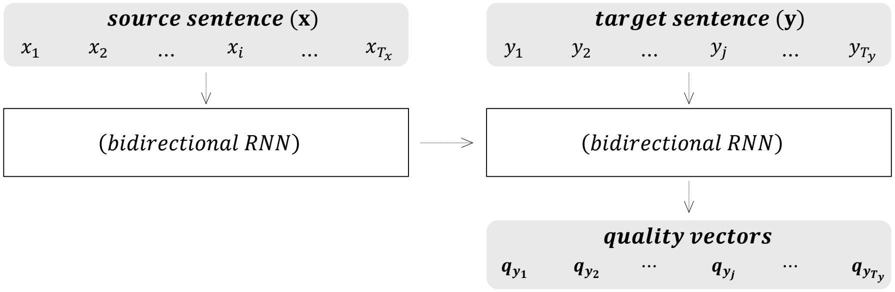
source sentence (x)
x_1 x_2 ... x_i ... x_{T_x}
target sentence (y)
y_1 y_2 ... y_j ... y_{T_y}
(bidirectional RNN)
(bidirectional RNN)
quality vectors
q_{y_1} q_{y_2} ... q_{y_j} ... $q_{y_{T_y}}$

Figure 1: First part of recurrent neural network based quality estimation model for generating quality vectors (Kim and Lee, 2016)

sentences as inputs,[2] 2) apply bidirectional RNNs both on source and target sentences, which enable to fully utilize the bidirectional quality information, and 3) generate quality vectors for target words as outputs.

In the second part (Figure 2, 3 and 4), the final quality of translation at various level (sentence-level/word-level/phrase-level) is predicted by using the quality vectors as inputs. Kim and Lee (2016) apply RNN based model to sentence-level QE and we extend this model to word and phrase-level QE. In subsection 2.1, 2.2 and 2.3, we describe the RNN based[3] (second part) sentence, word and phrase-level QE models.[4]

The cause of these separated parts of the QE model comes from the insufficiency of QE datasets to train the whole QE model. Thus, the QE model is divided into two parts, and then different training data are used to train each of the separated parts: large-scale parallel corpora such as Europarl for training the first part and QE datasets, provided in Quality Estimation Shared Task of WMT, for training the second part.

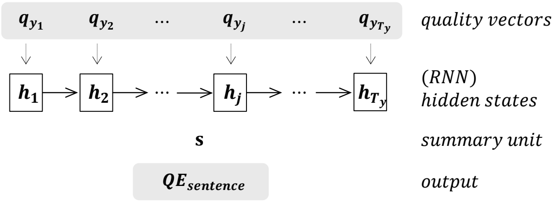
q_{y_1} q_{y_2} ... q_{y_j} ... $q_{y_{T_y}}$
quality vectors
h_1 h_2 ... h_j ... h_{T_y}
(RNN)
hidden states
s
summary unit
$QE_{sentence}$
output

Figure 2: Recurrent neural network based sentence-level QE model (SENT/RNN) (Kim and Lee, 2016)

2.1 RNN based Sentence-level QE Model

In RNN based sentence-level QE model (Figure 2), HTER (human-targeted translation edit rate) (Snover et al., 2006) in [0,1] for target sentence is predicted by using a logistic sigmoid function such that

$$\begin{aligned} &\mathrm{QE}_{sentence}(\, \mathbf{y},\, \mathbf{x}\,) \\ &\quad = \mathrm{QE}'_{sentence}(\, q_{y_1},\, \ldots\,,\, q_{y_{T_\mathbf{y}}}\,) \\ &\quad = \sigma(W_s\, \mathbf{s})\,. \end{aligned} \tag{1}$$

W_s is the weight matrix of sigmoid function[5] at sentence-level QE. $\mathbf{s}$ is a summary unit of the sequential quality vectors and is fixed to the last hidden state[6] $h_{T_\mathbf{y}}$ of RNN. The hidden state h_j is computed by

$$h_j = f(q_{y_j}, h_{j-1}) \tag{2}$$

where f is the activation function of RNN (Kim and Lee, 2016).

2.2 RNN based Word-level QE Model

In RNN based word-level QE model (Figure 3), we apply bidirectional RNN based binary classification (OK/BAD) using quality vectors as inputs. Through the bidirectional RNN, bidirectional hidden states $\{\vec{h}_j,\, \overleftarrow{h}_j\}$ for each target word y_j are

[2] In MT/NMT, only source sentence is used as a input. In QE, however, both source and target sentences can be used as inputs.

[3] In all activation functions of RNN, the gated hidden unit (Cho et al., 2014) is used to learn long-term dependencies.

[4] We, also, apply feedforward neural network (FNN) to the second part of QE model (see Appendix A). However, to reflect the dependencies between quality vectors and to fully utilize QE related information from QE datasets, we focus on the RNN based model.

[5] Bias terms are visually omitted in all equations.

[6] In RNN, the last hidden state is used as the summary of inputs.

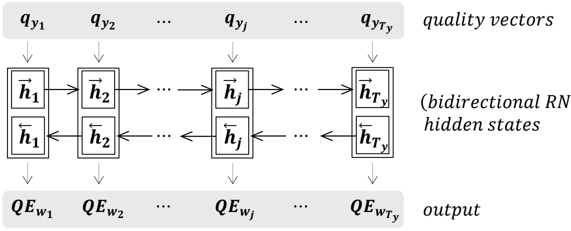
q_{y_1} q_{y_2} ··· q_{y_j} ··· $q_{y_{T_y}}$
quality vectors
$\vec{h}_1$ $\vec{h}_2$ ··· $\vec{h}_j$ ··· $\vec{h}_{T_y}$
$\overleftarrow{h}_1$ $\overleftarrow{h}_2$ ··· $\overleftarrow{h}_j$ ··· $\overleftarrow{h}_{T_y}$
(bidirectional RNN) hidden states
QE_{w_1} QE_{w_2} ··· QE_{w_j} ··· $QE_{w_{T_y}}$
output

Figure 3: Recurrent neural network based word-level QE model (WORD/RNN)

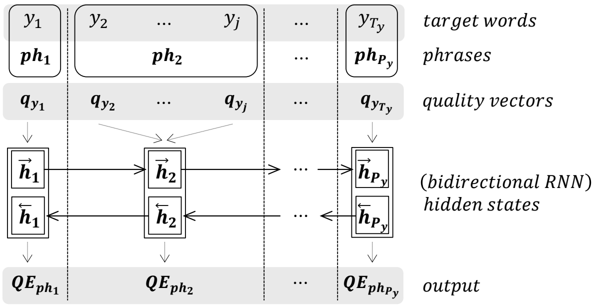
y_1 y_2 ··· y_j ··· y_{T_y}
target words
ph_1 ph_2 ··· ph_{P_y}
phrases
q_{y_1} q_{y_2} ··· q_{y_j} ··· $q_{y_{T_y}}$
quality vectors
$\vec{h}_1$ $\vec{h}_2$ ··· $\vec{h}_{P_y}$
$\overleftarrow{h}_1$ $\overleftarrow{h}_2$ ··· $\overleftarrow{h}_{P_y}$
(bidirectional RNN) hidden states
QE_{ph_1} QE_{ph_2} ··· $QE_{ph_{P_y}}$
output

Figure 4: Recurrent neural network based phrase-level QE model (PHR/RNN)

made such that[7]

$$\begin{aligned} \vec{h}_j &= f'(q_{y_j}, \vec{h}_{j-1}) \\ \overleftarrow{h}_j &= g'(q_{y_j}, \overleftarrow{h}_{j+1})\,. \end{aligned} \tag{3}$$

The forward hidden state $\vec{h}_j$ indicates summary information about the forward translation quality of target word y_j, reflecting qualities of preceding target words $\{y_1, ..., y_{j-1}\}$. And the backward hidden state $\overleftarrow{h}_j$ indicates summary information about the backward translation quality of target word y_j, reflecting qualities of following target words[8] $\{y_{j+1}, ..., y_{T_\mathbf{y}}\}$. We use the concatenated hidden state h_j^w $(= [\vec{h}_j; \overleftarrow{h}_j])$ to predict the word-level quality for target word y_j such that

$$\begin{aligned} &\mathrm{QE}_{word}(\, y_j,\, \mathbf{x}\,) \\ &\quad = \mathrm{QE}'_{word}(\, q_{y_j}\,) \\ &\quad = \begin{cases} \mathrm{OK}\ \ , & \text{if } \sigma(\, W_w\, h_j^w\,) \geq 0.5 \\ \mathrm{BAD}, & \text{if } \sigma(\, W_w\, h_j^w\,) < 0.5\,. \end{cases} \end{aligned} \tag{4}$$

W_w is the weight matrix of sigmoid function at word-level QE.

2.3 RNN based Phrase-level QE Model

RNN based phrase-level QE model is the extended version of RNN based word-level QE model (in subsection 2.2). In RNN based phrase-level QE model (Figure 4), we also apply bidirectional RNN based binary classification. We use the simply averaged quality vector q_{ph_j} to predict the phrase-level quality of the phrase[9] ph_j, composed of the corresponding target words $\{y_k, y_{k+1}, ...\}$, such that

$$\begin{aligned} &\mathrm{QE}_{phrase}(\, ph_j,\, \mathbf{x}\,) \\ &\quad = \mathrm{QE}'_{phrase}(\, q_{y_k}, q_{y_{k+1}}, \dots\,) \\ &\quad = \mathrm{QE}''_{phrase}(\, q_{ph_j}\,) \\ &\quad = \begin{cases} \mathrm{OK}\ \ , & \text{if } \sigma(\, W_{ph}\, h_j^{ph}\,) \geq 0.5 \\ \mathrm{BAD}, & \text{if } \sigma(\, W_{ph}\, h_j^{ph}\,) < 0.5\,. \end{cases} \end{aligned} \tag{5}$$

W_{ph} is the weight matrix of sigmoid function at phrase-level QE. h_j^{ph} $(= [\vec{h}_j; \overleftarrow{h}_j])$ is the concatenated hidden state for phrase ph_j of bidirectional RNN where[10]

$$\begin{aligned} \vec{h}_j &= f''(q_{ph_j}, \vec{h}_{j-1}) \\ \overleftarrow{h}_j &= g''(q_{ph_j}, \overleftarrow{h}_{j+1})\,. \end{aligned} \tag{6}$$

3 Results

RNN based QE models were evaluated on the WMT16 Quality Estimation Shared Task[11] at sentence, word and phrase level of English-German. Because whole QE models are separated into two parts, each part of the QE models is trained separately by using different training data. To train the first part of the QE models, English-German parallel corpus of Europarl v7 (Koehn, 2005) were used. To train the second part of the QE models, WMT16 QE datasets of English-German (Specia et al., 2016) were used.

To denote the each method, the following naming format is used: $[level]/[model]$-QV$[num]$. $[level]$ is the QE granularity level: SENT (sentence level), WORD (word level) and PHR (phrase level). $[model]$ is the type of model used in the second part: RNN (of subsection 2.1, 2.2 and 2.3)

[7] $f'(g')$ is the activation function of the forward(backward) RNN at word-level QE.

[8] $T_\mathbf{y}$ is the length of target sentence.

[9] $1 \leqq j \leqq P_\mathbf{y}$ where $P_\mathbf{y}$ is the number of phrases in target sentence and $P_\mathbf{y} \leq T_\mathbf{y}$.

[10] $f''(g'')$ is the activation function of the forward(backward) RNN at phrase-level QE.

[11] http://www.statmt.org/wmt16/quality-estimation-task.html

Task 1. Test	**Pearson's** $r \uparrow$	MAE $\downarrow$	RMSE $\downarrow$	Rank
SENT/RNN-QV2	0.4600	0.1358	0.1860	2
SENT/RNN-QV3	0.4475	0.1352	0.1838	4
SENT/FNN-QV2	0.3588	0.1517	0.2001	
SENT/FNN-QV3	0.3549	0.1529	0.2006	
BASELINE	0.3510	0.1353	0.1839	

Table 1: Results on **test set** for the **scoring** variant of WMT16 **sentence-level** QE (Task 1).

Task 1. Test	**Spearman's** $\rho \uparrow$	DeltaAvg $\uparrow$	Rank
SENT/RNN-QV2	0.4826	0.0766	1
SENT/RNN-QV3	0.4660	0.0753	3
SENT/FNN-QV3	0.3910	0.0589	
SENT/FNN-QV2	0.3905	0.0593	
BASELINE	0.3900	0.0630	

Table 2: Results on **test set** for the **ranking** variant of WMT16 **sentence-level** QE (Task 1).

Task 2. Test	**Multiplication of**	F1-	F1-	
Word-level	**F1-OK and F1-BAD** $\uparrow$	Bad $\uparrow$	OK $\uparrow$	Rank
WORD/RNN-QV3	0.3803	0.4475	0.8498	5
WORD/RNN-QV2	0.3759	0.4538	0.8284	6
WORD/FNN-QV3	0.3273	0.3800	0.8615	
WORD/FNN-QV2	0.3241	0.3932	0.8242	
BASELINE	0.3240	0.3682	0.8800	

Table 3: Results on **test set** of WMT16 **word-level** QE (Task 2).

Task 2. Test	**Multiplication of**	F1-	F1-	
Phase-level	**F1-OK and F1-BAD** $\uparrow$	Bad $\uparrow$	OK $\uparrow$	Rank
PHR/RNN-QV3	0.3781	0.4950	0.7639	2
PHR/RNN-QV2	0.3693	0.4785	0.7718	3
PHR/FNN-QV3	0.3505	0.4722	0.7423	
PHR/FNN-QV2	0.3353	0.4413	0.7599	
BASELINE	0.3211	0.4014	0.8001	

Table 4: Results on **test set** of WMT16 **phrase-level** QE (Task 2).

and FNN (of subsection A.1, A.2 and A.3). At QV$[num]$, $[num]$ is the number of iterations while the first part is trained by using large-scale parallel corpora to make quality vectors (QV).

3.1 Results of Sentence-level QE (Task 1)

Pearson's correlation (r), mean absolute error (MAE), and root mean squared error (RMSE) are used to evaluate the scoring variant of sentence-level QE. And Spearman's rank correlation (ρ) and DeltaAvg are used to evaluate the ranking variant of sentence-level QE.

Table 1 and 2 (Table B.1 and B.2) present the results of the QE models on test (development) set for the scoring and ranking variants of the WMT16 sentence-level QE shared task (Task 1). In all aspects of evaluation at sentence-level QE, the RNN based QE model (SENT/RNN) showed the better performance than the FNN based QE model (SENT/FNN). Our two methods (SENT/RNN-QV2 and SENT/RNN-QV3), participated in WMT16 sentence-level QE shared task, achieved top rank: each 2nd and 4th at the scoring variant and each 1st and 3rd at the ranking variant.

3.2 Results of Word-level and Phrase-level QE (Task 2)

The multiplication of F1-scores for the 'OK' and 'BAD' classes and F1-score for the 'BAD' class are used to evaluate the word-level and phrase-level QE.

Table 3 and 4 (Table B.3 and B.4) respectively present the results on test (development) set of the WMT16 word-level and phrase-level QE shared task (Task 2). In all aspects of evaluation at word-level and phrase-level QE, the RNN based QE models (WORD/RNN and PHR/RNN) showed the better performance than the FNN based QE models (WORD/FNN and PHR/FNN). Our two methods (WORD/RNN-QV3 and WORD/RNN-QV2), participated in WMT16 word-level QE shared task, achieved each 5th and 6th rank. Our two methods (PHR/RNN-QV3 and PHR/RNN-QV2), participated in WMT16 phrase-level QE shared task, achieved top rank: each 2nd and 3rd.

4 Conclusion

This paper described recurrent neural network based quality estimation models of sentence, word and phrase level. We extended the (existing sentence-level) recurrent neural network based quality estimation model to word and phrase level. And we applied these models to sentence, word and phrase-level QE shared task of WMT16. These recurrent neural network based quality estimation models are pure neural approaches for QE and achieved excellent performance especially in sentence and phrase-level QE.

Acknowledgments

This research was partly supported by the "ICT R&D Program" of MSIP/IITP (R7119-16-1001, Core technology development of the real-time simultaneous speech translation based on knowledge enhancement) and "ICT Consilience Creative Program" of MSIP/IITP (R0346-16-1007)

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics.

Hyun Kim and Jong-Hyeok Lee. 2016. A recurrent neural networks approach for estimating the quality of machine translation output. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 494–498. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Julia Kreutzer, Shigehiko Schamoni, and Stefan Riezler. 2015. Quality estimation from scratch (quetch): Deep learning for word-level translation quality estimation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 316–322. Association for Computational Linguistics.

Kashif Shah, Varvara Logacheva, Gustavo Paetzold, Frédéric Blain, Daniel Beck, Fethi Bougares, and Lucia Specia. 2015. Shef-nn: Translation quality estimation with neural networks. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 342–347. Association for Computational Linguistics, September.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with quest++. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 115–120. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

Lucia Specia, Varvara Logacheva, and Carolina Scarton. 2016. WMT16 quality estimation shared task training and development data. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Appendix

A Feedforward Neural Network (FNN) based Quality Estimation Model

In this section, we describe the FNN based (second part) QE models of sentence, word and phrase level, for comparison with RNN based (second part) QE models. Quality vectors, generated from the same RNN based first part QE model, are used as inputs.

A.1 FNN based Sentence-level QE Model

In FNN based sentence-level QE model (Figure A.1), we also use a logistic sigmoid function of (1). But in FNN based model we make each hidden state h_j by only using each quality vector q_{y_j} for target word y_j. And for $\mathbf{s}$, which is a summary unit of the whole quality vectors, we simply average all hidden states $\{h_1, ..., h_{T_\mathbf{y}}\}$.

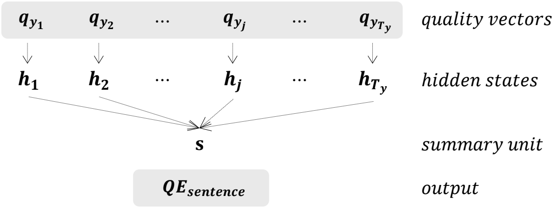

Figure A.1: Feedforward neural network based sentence-level QE model (SENT/FNN)

A.2 FNN based Word-level QE Model

In FNN based word-level QE model (Figure A.2), we apply FNN based binary classification (OK/BAD) using quality vectors as input. By only using each quality vector q_{y_j} for target word y_j, each hidden state $h_j (= h_j^w)$ is made. We predict the word-level QE by (4).

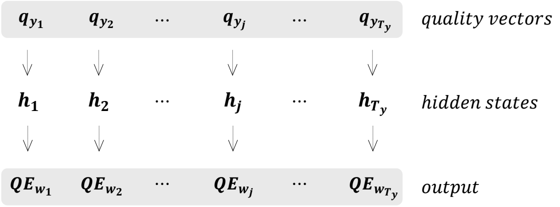

Figure A.2: Feedforward neural network based word-level QE model (WORD/FNN)

A.3 FNN based Phrase-level QE Model

In FNN based phrase-level QE model (Figure A.3), we also apply FNN based binary classification (OK/BAD). By only using the averaged quality vector q_{ph_j} for the phrase ph_j, composed of the corresponding target words $\{y_k, y_{k+1}, ...\}$, the hidden state $h_j (= h_j^{ph})$ is made. We predict the phrase-level QE by (5).

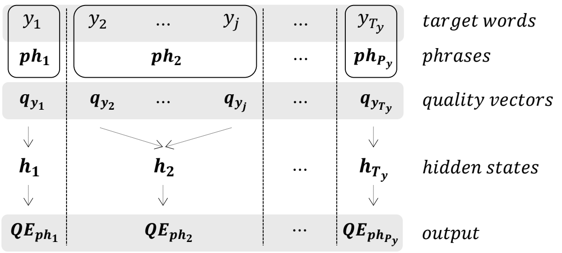

Figure A.3: Feedforward neural network based phrase-level QE model (PHR/FNN)

B Results on Development Set of WMT16 QE Shared Task

Task 1. Dev	**Pearson's** $r \uparrow$	MAE $\downarrow$	RMSE $\downarrow$
SENT/RNN-QV2	0.4661	0.1340	0.1921
SENT/RNN-QV3	0.4658	0.1341	0.1896
SENT/FNN-QV3	0.3915	0.1539	0.2007
SENT/FNN-QV2	0.3904	0.1560	0.2015

Table B.1: Results on **development set** for the **scoring** variant of WMT16 **sentence-level** QE (Task 1).

Task 1. Dev	**Spearman's** $\rho \uparrow$	DeltaAvg $\uparrow$
SENT/RNN-QV3	0.5222	0.0882
SENT/RNN-QV2	0.5154	0.0892
SENT/FNN-QV3	0.4370	0.0697
SENT/FNN-QV2	0.4227	0.0693

Table B.2: Results on **development set** for the **ranking** variant of WMT16 **sentence-level** QE (Task 1).

Task 2. Dev. Word-level	**Multiplication of** **F1-OK and F1-BAD** $\uparrow$	F1- Bad $\uparrow$	F1- OK $\uparrow$
WORD/RNN-QV3	0.3880	0.4567	0.8496
WORD/RNN-QV2	0.3838	0.4617	0.8313
WORD/FNN-QV3	0.3227	0.3812	0.8597
WORD/FNN-QV2	0.3171	0.3878	0.8178

Table B.3: Results on **development set** of WMT16 **word-level** QE (Task 2).

Task 2. Dev. Phase-level	**Multiplication of** **F1-OK and F1-BAD** $\uparrow$	F1- Bad $\uparrow$	F1- OK $\uparrow$
PHR/RNN-QV2	0.3831	0.4955	0.7731
PHR/RNN-QV3	0.3770	0.4975	0.7578
PHR/FNN-QV3	0.3526	0.4755	0.7416
PHR/FNN-QV2	0.3391	0.4447	0.7626

Table B.4: Results on **development set** of WMT16 **phrase-level** QE (Task 2).