

Multimodal quality estimation for machine translation

Item Type	Conference contribution
Authors	Okabe, Shu;Blain, Frédéric;Specia, Lucia
Citation	Okabe, S., Blain, F. and Specia, L. (2020) Multimodal quality estimation for machine translation. In, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Jurafsky, D., Chai, J., Schluter, N. and Tetreault, J. (eds.) Stroudsburg, PA: Association for Computational Linguistics. pp. 1233-1240
DOI	10.18653/v1/2020.acl-main.114
Publisher	Association for Computational Linguistics
Download date	2026-04-16 05:12:40
License	http://creativecommons.org/licenses/by/4.0/
Link to Item	http://hdl.handle.net/2436/623554

Multimodal Quality Estimation for Machine Translation

Shu Okabe*¹

Frédéric Blain*²

Lucia Specia^{1,2}

¹Department of Computing, Imperial College London, UK

²Department of Computer Science, University of Sheffield, UK

shurokabe@gmail.com

f.blain@sheffield.ac.uk

l.specia@imperial.ac.uk

Abstract

We propose approaches to Quality Estimation (QE) for Machine Translation that explore both text and visual modalities for Multimodal QE. We compare various multimodality integration and fusion strategies. For both sentence-level and document-level predictions, we show that state-of-the-art neural and feature-based QE frameworks obtain better results when using the additional modality.

1 Introduction

Quality Estimation (QE) for Machine Translation (MT) (Blatz et al., 2004; Specia et al., 2009) aims to predict the quality of a machine-translated text without using reference translations. It estimates a label (a category, such as ‘good’ or ‘bad’, or a numerical score) for a translation, given text in a source language and its machine translation in a target language (Specia et al., 2018b). QE can operate at different linguistic levels, including sentence and document levels. Sentence-level QE estimates the translation quality of a whole sentence, while document-level QE predicts the translation quality of an entire document, even though in practice in literature the documents have been limited to a small set of 3-5 sentences (Specia et al., 2018b).

Existing work has only explored textual context. We posit that to judge (or estimate) the quality of a translated text, additional context is paramount. Sentences or short documents taken out of context may lack information on the correct translation of certain (esp. ambiguous) constructions. Inspired by recent work on multimodal machine learning (Baltrusaitis et al., 2019; Barrault et al., 2018), we propose to explore the visual modality in addition to the text modality for this task.

Multimodality through vision offers interesting opportunities for real-life data since texts are in-


Source (EN)	Danskin Women’s Bermuda Shorts	
MT (FR)	Bermuda Danskin féminines court	

Table 1: Example of incorrectly machine-translated text: the word *shorts* is used to indicate short trousers, but gets translated in French as *court*, the adjective *short*. Here multimodality could help to detect the error (extracted from the Amazon Reviews Dataset of McAuley et al., 2015).

creasingly accompanied with visual elements such as images or videos, especially in social media but also in domains such as e-commerce. Multimodality has not yet been applied to QE. Table 1 shows an example from our e-commerce dataset in which multimodality could help to improve QE. Here, the English noun *shorts* is translated by the adjective *court* (for the adjective *short*) in French, which is a possible translation out of context. However, as the corresponding product image shows, this product is an item of clothing, and thus the machine translation is incorrect. External information can hence help identify mismatches between translations which are difficult to find within the text.

Progress in QE is mostly benchmarked as part of the Conference on Machine Translation (WMT) Shared Task on QE. This paper is based on data from the WMT’18 edition’s Task 4 – document-level QE. This Task 4 aims to predict a translation quality score for short documents based on the number and the severity of translation errors at the word level (Specia et al., 2018a). This data was chosen as it is the only one for which meta information (images in this case) is available. We extend this dataset by computing scores for each sentence for a sentence-level prediction task. We consider both feature-based and neural state-of-the-art models for QE. Having these as our starting

*Two authors contributed equally.

points, we propose different ways to integrate the visual modality.

The main contributions of this paper are as follows: (i) we introduce the task of Multimodal QE (MQE) for MT as an attempt to improve QE by using external sources of information, namely images; (ii) we propose several ways of incorporating visual information in neural-based and feature-based QE architectures; and (iii) we achieve the state-of-the-art performance for such architectures in document and sentence-level QE.

2 Experimental Settings

2.1 QE Frameworks and Models

We explore feature-based and neural-based models from two open-source frameworks:

QuEst++: QuEst++ (Specia et al., 2015) is a feature-based QE framework composed of two modules: a feature extractor module, to extract the relevant QE features from both the source sentences and their translations, and a machine learning module. We only use this framework for our experiments on document-level QE, since it does not perform well enough for sentence-level prediction. We use the same model (Support Vector Regression), hyperparameters and feature settings as the baseline model for the document-level QE task at WMT’18.

deepQuest: deepQuest (Ive et al., 2018) is a neural-based framework that provides state-of-the-art models for multi-level QE. We use the BiRNN model, a light-weight architecture which can be trained at either sentence or document level.

The BiRNN model uses an encoder-decoder architecture: it takes on its input both the source sentence and its translation which are encoded separately by two independent bi-directional Recurrent Neural Networks (RNNs). The two resulting sentence representations are then concatenated as a weighted sum of their word vectors, generated by an attention mechanism. For sentence-level predictions, the weighted representation of the two input sentences is passed through a dense layer with sigmoid activation to generate the quality estimates. For document-level predictions, the final representation of a document is generated by a second attention mechanism, as the weighted sum of the weighted sentence-level representations of all the sentences within the document. The resulting document-level representation is then passed

through a dense layer with sigmoid activation to generate the quality estimates.

Additionally, we propose and experiment with BERT-BiRNN, a variant of the BiRNN model. Rather than training the token embeddings with the task at hand, we use large-scale pre-trained token-level representations from the multilingual cased base BERT model (Devlin et al., 2019). During training, the BERT model is fine-tuned by unfreezing the weights of the last four hidden layers along with the token embedding layer. This performs comparably to the state-of-the-art predictor-estimator neural model in Kepler et al. (2019).

2.2 Data

WMT’18 QE Task 4 data: This dataset was created for the document-level track. It contains a sample of products from the Amazon Reviews Dataset (McAuley et al., 2015) taken from the Sports & Outdoors category. ‘Documents’ consist of the English product title and its description, its French machine-translation and a numerical score to predict, namely the MQM score (Multidimensional Quality Metrics) (Lommel et al., 2014). This score is computed by annotating and weighting each word-level translation error according to its severity (minor, major and critical):

$$\text{MQM Score} = 1 - \frac{n_{min} + 5n_{maj} + 10n_{cri}}{n}$$

where n is the total number of words, and n_i is the number of errors annotated with the corresponding error severity. Additionally, the dataset provides one picture per product, as well as pre-extracted visual features, as we discuss below.

For the sentence-level QE task, each document of the dataset was split into sentences (lines), where every sentence has its corresponding MQM score computed in the same way as for the document. We note that this variant is different from the official sentence-level track at WMT since for that task visual information is not available.

Text features: For the feature-based approach, we extract the same 15 features as those for the baseline of WMT’18 at document level. For the neural-based approaches, text features are either the learned word embeddings (BiRNN) or pre-trained word embeddings (BERT-BiRNN).

Visual features: The visual features are pre-extracted vectors with 4,096 dimensions, also provided in the Amazon Reviews Dataset (McAuley

et al., 2015). The method to obtain the features uses a deep convolutional neural network which has been pre-trained on the ImageNet dataset for image classification (Deng et al., 2009). The visual features extracted represent a vectorial summary of the image taken from the last pooled layer of the network. He and McAuley (2016) have shown that this representation contains useful visual features for a number of tasks.

3 Multimodal QE

We propose different ways to integrate visual features in our two monomodal QE approaches (Sections 3.1 and 3.2). We compare each proposed model with its monomodal QE counterpart as baseline, both using the same hyperparameters.

3.1 Multimodal feature-based QE

The feature-based textual features contain 15 numerical scores, while the visual feature vector contains 4,096 dimensions. To avoid over-weighting the visual features, we reduce their dimensionality using Principal Component Analysis (PCA). We consider up to 15 principal components in order to keep a balance between the visual features and the 15 text features from QuEst++. We choose the final number of principal components to keep according to the explained variance with the PCA, so this number is treated as a hyperparameter. After analysing the explained variance for up to 15 kept principal components (see Figure 4 in Appendix), we selected six numbers of principal components to train QE models with (1, 2, 3, 5, 10, and 15). As fusion strategy, we concatenate the two feature vectors.

3.2 Multimodal neural-based QE

Multimodality is achieved with two changes in our monomodal models: **multimodality integration** (*where* to integrate the visual features in the architecture), and **fusion strategy** (*how* to fuse the visual and textual features). We propose the following places to integrate the visual feature vector into the BiRNN architecture:

- **embed** – the visual feature vector is used after the word embedding layer;
- **annot** – the visual feature vector is used after the encoding of the two input sentences by the two bi-directional RNNs;

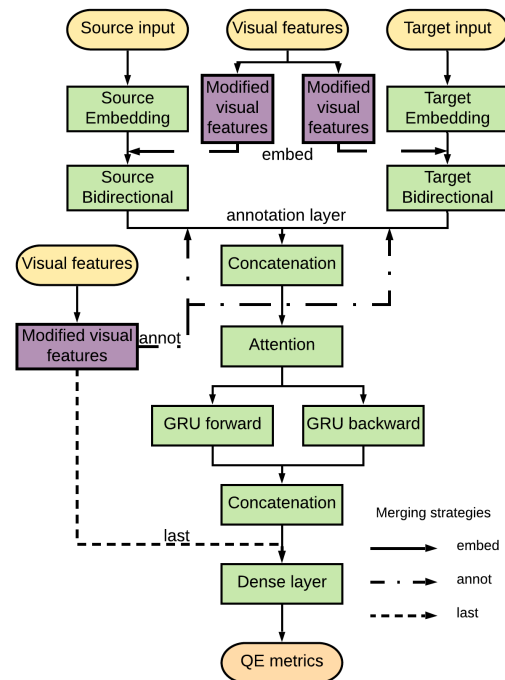


Figure 1: High-level representation of the document-level BiRNN architecture which illustrates how the visual features are integrated into the model. The three different strategies are ‘embed’, ‘annot’ and ‘last’.

- **last** – the visual feature vector is used just before the last layer.

To fuse the visual and text features, we reduce the size of the visual features using a dense layer with a ReLU activation and reshape it to match the shape of the text-feature vector. As fusion strategies between visual and textual feature vectors, we propose the following:

- **conc** – concatenation with both source and target word representations for the ‘embed’ strategy; concatenation with the text features for the ‘last’ strategy;
- **mult** – element-wise multiplication for the target word representations and concatenation for the source word representations for the ‘embed’ strategy; element-wise multiplication with the text features for the ‘annot’ and ‘last’ strategies;
- **mult2** – element-wise multiplication for both source and target word representations (exclusive to the ‘embed’ model).

Figure 1 presents the high-level architecture of the document-level BiRNN model, with the various multimodality integration and fusion approaches.

For example, in the ‘embed’ setting, the visual features are fused with each word representation from the embedding layers. Since this strategy modifies the embedding for each word, it can be expected to have a bigger impact on the result.

4 Results

We use the standard training, development and test datasets from the WMT’18 Task 4 track. For feature-based systems, we follow the built-in cross-validation in QuEst++, and train a single model with the hyperparameters found by cross-validation. For neural-based models, we use early-stopping with a patience of 10 to avoid over-fitting, and all reported figures are averaged over 5 runs corresponding to different seeds.

We follow the evaluation method of the WMT QE tasks: Pearson’s r correlation as the main metric (Graham, 2015), Mean-Absolute Error (MAE) and Root-Mean-Squared Error (RMSE) as secondary metrics. For statistical significance on Pearson’s r , we compute Williams test (Williams, 1959) as suggested by Graham and Baldwin (2014).

For all neural-based models, we experiment with the all three integration strategies (‘embed’, ‘annot’ and ‘last’) and all three fusion strategies (‘conc’, ‘mult’ and ‘mult2’) presented in Section 3.2. This leads to 6 multimodal models for each BiRNN and BERT-BiRNN. In Tables 2 and 4, as well as in Figures 2 and 3, we report the top three performing models. We refer the reader to the Appendix for the full set of results.

4.1 Sentence-level MQE

The first part of Table 2 presents the results for sentence-level multimodal QE with BiRNN. The best model is BiRNN+Vis-embed-mult2, achieving a Pearson’s r of 0.535, significantly outperforming the baseline (p -value<0.01). Visual features can, therefore, help to improve the performance of sentence-level neural-based QE systems significantly.

Figure 2 presents the result of Williams significance test for BiRNN model variants. It is a correlation matrix that can be read as follows: the value in cell (i, j) is the p -value of Williams test for the change in performance of the model at row i compared to the model at column j (Graham, 2015).

With the pre-trained token-level representations from BERT (second half of Table 2), the best model is BERT-BiRNN+Vis-annot-mult, achieving a Pear-

	Pearson	MAE	RMSE
BiRNN	0.504	0.539	0.754
+Vis-last-conc	0.483	0.531	0.746
+Vis-embed-mult	0.473	0.534	0.753
+Vis-embed-mult2	0.535	0.569	0.792
BERT-BiRNN	0.590	0.455	0.659
+Vis-annot-mult	0.602	0.454	0.654
+Vis-embed-conc	0.576	0.474	0.694
+Vis-embed-mult	0.598	0.486	0.686

Table 2: Pearson correlation at **sentence-level** on the WMT’18 dataset. We report the monomodal models (BiRNN, BERT-BiRNN) and their respective top-3 best performing multimodal variants (+Vis). We refer the reader to the Appendix for the full set of results.

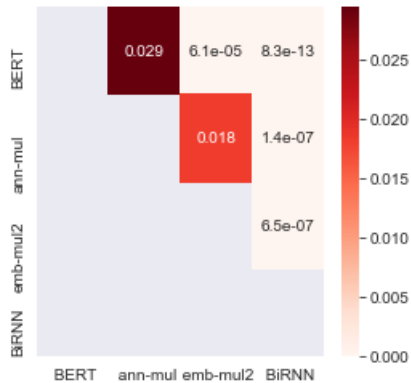


Figure 2: Williams significance test of top models for **sentence-level** BiRNN on the WMT’18 dataset. Here, *BERT*, *ann-mul* and *emb-mul2* correspond to the BERT-BiRNN, the BERT-BiRNN+Vis-annot-mult and the BiRNN+Vis-embed-mult2 models of Table 2.

son’s r of 0.602. This shows that even when using better word presentations, the visual features help to get further (albeit modest) improvements.

Table 3 shows an example of predicted scores at the sentence-level for the baseline model (BiRNN) and for the best multimodal BiRNN model (BiRNN+Vis-embed-mult2). The multimodal model has predicted a closer score (-0.002) to the gold MQM score (0.167) than the baseline model (-0.248). The French translation is poor (*cumulative-split* is, for instance, not translated) as the low gold MQM score shows. However, the (main) word *stopwatch* is correctly translated as *chronomètre* in French. Since the associated picture indeed represents a stopwatch, one explanation for this improvement could be that the multimodal model may have rewarded this correct and important part of the translation.


Source (EN)	The A601X stopwatch features cumulative-split timing.	
MT (FR)	Le chronomètre A601X dispose calendrier cumulative-split.	
gold MQM score	0.167	
BiRNN	-0.248	
BiRNN+Vis-embed-mult2	-0.002	

Table 3: Example of performance of **sentence-level** multimodal QE. Compared to the baseline prediction (BiRNN), the prediction from the best multimodal model (BiRNN+Vis-embed-mult2) is closer to the gold MQM score. This could be because the word *stopwatch* is correctly translated as *chronomètre* in French, and the additional visual feature confirms it. This could lead to an increase in the predicted score to reward the correct part, despite the poor translation (extracted from the Amazon Reviews Dataset of McAuley et al., 2015).

4.2 Document-level MQE

Table 4 presents the results for the document-level feature-based and BiRNN neural QE models.¹ The first section shows the official models from the WMT’18 QE Task 4 report (Specia et al., 2018a). The neural-based approach SHEF-PT is the winning submission, outperforming another neural-based approach (SHEF-mtl-bRNN). For our BiRNN models (second section), BiRNN+Vis-embed-conc performs only slightly better than the monomodal baseline. For the feature-based models (third section), on the other hand, the baseline monomodal QuEst++ is outperformed by various multimodal variants by a large margin, with the one with two principal components (QuEst+Vis-2) performing the best. The more PCA components kept, the worse the results (see Appendix for full set of results).

	Pearson	MAE	RMSE
SHEF-PT	0.534	0.562	0.852
SHEF-mtl-bRNN	0.473	0.566	–
BiRNN	0.495	0.531	0.788
+Vis-annot-mult	0.494	0.531	0.793
+Vis-embed-conc	0.501	0.536	0.780
+Vis-embed-mult2	0.491	0.575	0.831
QuEst	0.503	0.547	0.802
+Vis-2	0.536	0.534	0.791
+Vis-3	0.528	0.538	0.793
+Vis-5	0.520	0.539	0.797

Table 4: Pearson correlation at **document-level** on the WMT’18 dataset: state-of-the-art models as reported by task organisers, our BiRNN model and its multimodal versions and feature-based QuEst++ and its multimodal versions.

Figure 3 shows the Williams significance test for document-level QuEst++ on the WMT’18 dataset.

¹The BERT-BiRNN models performed very poorly at this level and more research on why is left for future work.

As we can see, QuEst+Vis-2 model outperforms the baseline with p -value = 0.002. Thus, visual features significantly improve the performance of feature-based QE systems compared to the monomodal QE counterparts.

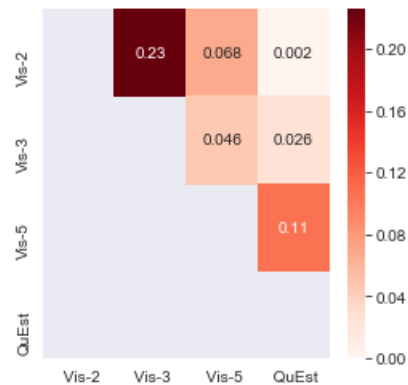


Figure 3: Williams significance test of top models for **document-level** QuEst++ on the WMT’18 dataset.

5 Conclusions

We introduced Multimodal Quality Estimation for Machine Translation, where an external modality – visual information – is incorporated to feature-based and neural-based QE approaches, on sentence and document levels. The use of visual features extracted from images has led to significant improvements in the results of state-of-the-art QE approaches, especially at sentence level.

The version of deepQuest for multimodal QE and scripts to convert document into sentence-level data are available on <https://github.com/sheffieldnlp/deepQuest>.

Acknowledgments

This work was supported by funding from both the Bergamot project (EU H2020 Grant No. 825303) and the MultiMT project (EU H2020 ERC Starting Grant No. 678017).

References

- Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. **Multimodal machine learning: A survey and taxonomy**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.
- Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. **Findings of the third shared task on multimodal machine translation**. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 308–327, Belgium, Brussels. Association for Computational Linguistics.
- John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchez, and Nicola Ueffing. 2004. **Confidence estimation for machine translation**. In *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland.
- J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. **Imagenet: A large-scale hierarchical image database**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **BERT: Pre-training of deep bidirectional transformers for language understanding**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. **Findings of the wmt 2019 shared tasks on quality estimation**. In *Proceedings of the Fourth Conference on Machine Translation*, pages 230–239, Florence, Italy. Association for Computational Linguistics.
- Yvette Graham. 2015. **Improving evaluation of machine translation quality estimation**. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1804–1813, Beijing, China. Association for Computational Linguistics.
- Yvette Graham and Timothy Baldwin. 2014. **Testing for significance of increased correlation with human judgment**. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar. Association for Computational Linguistics.
- Ruining He and Julian McAuley. 2016. **Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering**. In *Proceedings of the 25th International Conference on World Wide Web*, pages 507–517, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.
- Julia Ive, Frédéric Blain, and Lucia Specia. 2018. **Deepquest: a framework for neural-based quality estimation**. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3146–3157.
- Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. **OpenKiwi: An open source framework for quality estimation**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics—System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.
- Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. **Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics**. *Tradumàtica: tecnologies de la traducció*, 0(12):455–463.
- Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. **Image-based recommendations on styles and substitutes**. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52, New York, NY, USA. ACM.
- Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón Astudillo, and André F. T. Martins. 2018a. **Findings of the WMT 2018 shared task on quality estimation**. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Belgium, Brussels. Association for Computational Linguistics.
- Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. **Estimating the sentence-level quality of machine translation systems**. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, pages 28–35, Barcelona, Spain.
- 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, Beijing, China. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.
- Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018b. **Quality Estimation for Machine Translation**. *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool.
- Evan J. Williams. 1959. *Regression Analysis*, volume 14. Wiley, New York, USA.

A Appendix

PCA analysis Figure 4 shows an almost linear relationship between the number of principal components and the explained variance of the PCA (see Section 3.1), i.e. the higher the number of principal components, the larger the explained variance. Therefore, we experimented with various numbers of components up to 15 (1, 2, 3, 5, 10, and 15) on the development set to find the best settings for quality prediction.

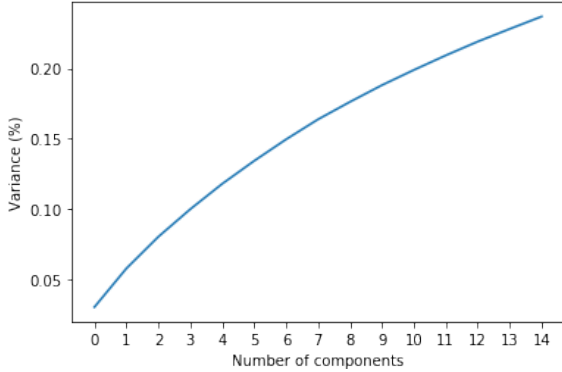


Figure 4: Explained variance of 15 components (cumulative sum) for the training set of the WMT’18 Task data at document level.

Complete results Tables 5 and 6 present the full set of results of our experiments on document and sentence-level multimodal QE on our main test set, the WMT’18 test set. These are a super-set of the results presented in the main paper but include all combinations of multimodality integration and fusion strategies for sentence-level prediction, as well as different numbers of principal components kept for document-level QuEst prediction models.

Additional test set Tables 7 and 8 present the full set of results of our experiments on the WMT’19 Task 2 test set on document and sentence-level multimodal QE, respectively. This was the follow-up edition of the WMT’18 Task 4, where the same training set is used, but a new test set is released.

For document-level, we observe nuanced results with more modest benefits in using visual features, regardless of the integration method or fusion strategy.

For sentence-level, we observe on the one hand quite significant improvements with a gain of almost 8 points in Pearson’s r over BiRNN, our monomodal baseline without pre-trained word embedding. It is interesting to note that almost all

	Pearson	MAE	RMSE
BiRNN	0.495	0.531	0.788
+Vis-last conc	0.476	0.550	0.802
+Vis-last-mult	0.481	0.543	0.812
+Vis-annot-mult	0.494	0.531	0.793
+Vis-embed-conc	0.501	0.536	0.780
+Vis-embed-mult	0.481	0.567	0.819
+Vis-embed-mult2	0.491	0.575	0.831
QuEst	0.503	0.547	0.802
+Vis-1	0.497	0.545	0.801
+Vis-2	0.536	0.534	0.790
+Vis-3	0.528	0.538	0.793
+Vis-5	0.520	0.539	0.797
+Vis-10	0.520	0.536	0.796
+Vis-15	0.515	0.540	0.801

Table 5: **Document-level** results for BiRNN and QuEst++ on the WMT’18 dataset, with and without visual features.

	Pearson	MAE	RMSE
BiRNN	0.504	0.539	0.754
+Vis-last-conc	0.483	0.531	0.746
+Vis-last-mult	0.462	0.511	0.733
+Vis-annot-mult	0.460	0.521	0.741
+Vis-embed-conc	0.467	0.541	0.765
+Vis-embed-mult	0.473	0.534	0.753
+Vis-embed-mult2	0.535	0.569	0.792
BERT-BiRNN	0.590	0.455	0.659
+Vis-last-conc	0.360	0.993	1.252
+Vis-last-mult	0.529	0.520	0.744
+Vis-annot-mult	0.602	0.454	0.654
+Vis-embed-conc	0.576	0.474	0.694
+Vis-embed-mult	0.598	0.486	0.686
+Vis-embed-mult2	0.570	0.573	0.770

Table 6: **Sentence-level** results for BiRNN and BERT-BiRNN on the WMT’18 Task 4 dataset, with and without visual features.

multimodal variants achieve better performance compared to the monomodal BiRNN baseline, with a peak when the visual features are fused with the word embedding representations by element-wise multiplication. On the other hand, we do not observe any gain in using visual features on the WMT’19 test set compared to our monomodal baseline with pre-trained word-embedding (BERT-BiRNN). Here that the BERT-BiRNN baseline model already performs very well. According to the task organisers, the mean MQM value on the WMT’19 test set is higher than on the WMT’18 test set, but actually closer to the training data (Fonseca

et al., 2019). We therefore hypothesise here that the highly dimensional and contextualised word-level representations from BERT are already enough and do not benefit from the extra information provided by the visual features.

	Pearson	MAE	RMSE
BiRNN	0.367	0.335	0.413
+Vis-last-conc	0.332	0.416	0.503
+Vis-last-mult	0.261	0.329	0.421
+Vis-annot-mult	0.332	0.276	0.353
+Vis-embed-conc	0.370	0.364	0.439
+Vis-embed-mult	0.335	0.313	0.398
+Vis-embed-mult2	0.344	0.285	0.361

Table 7: **Document-level** results for BiRNN on the WMT’19 Task 2 test set, with and without visual features.

Metrics	Pearson	MAE	RMSE
BiRNN	0.485	0.616	0.922
+Vis-last-conc	0.492	0.602	0.908
+Vis-last-mult	0.520	0.584	0.895
+Vis-annot-mult	0.508	0.591	0.901
+Vis-embed-conc	0.470	0.614	0.927
+Vis-embed-mult	0.474	0.613	0.927
+Vis-embed-mult2	0.563	0.609	0.944
BERT-BiRNN	0.652	0.556	0.842
+Vis-last-mult	0.605	0.568	0.854
+Vis-annot-mult	0.596	0.565	0.845
+Vis-embed-conc	0.594	0.571	0.853
+Vis-embed-mult	0.596	0.560	0.827
+Vis-embed-mult2	0.590	0.581	0.853

Table 8: **Sentence-level** results for BiRNN and BERT-BiRNN on the WMT’19 Task 2 test dataset, with and without visual features.