SimpleNets: Machine Translation Quality Estimation with Resource-Light Neural Networks

Gustavo Henrique Paetzold and **Lucia Specia**
Department of Computer Science
University of Sheffield, UK
{ghpaetzold1,l.specia}@sheffield.ac.uk

Abstract

We introduce SimpleNets: a resource-light solution to the sentence-level Quality Estimation task of WMT16 that combines Recurrent Neural Networks, word embedding models, and the principle of compositionality. The SimpleNets systems explore the idea that the quality of a translation can be derived from the quality of its n-grams. This approach has been successfully employed in Text Simplification quality assessment in the past. Our experiments show that, surprisingly, our models can learn more about a translation's quality by focusing on the original sentence, rather than on the translation itself.

1 Introduction

The task of Machine Translation Quality Estimation (QE) has gained noticeable popularity in the last few years. The goal of QE is to predict the quality of translations produced by a certain Machine Translation (MT) system in the absence of reference translations. Reliable solutions for QE can be useful in various tasks, such as improving post-editing efficiency (Specia, 2011), selecting high quality translations (Soricut and Echihabi, 2010), translation re-ranking (Shah and Specia, 2014), and visual assistance for manual translation revision (Bach et al., 2011).

QE can be performed in various ways in order to suit different purposes. The most widely addressed form of this task is sentence-level QE. Most existing work addresses this task as a supervised learning problem, in which a set of training examples is used to learn a model that predicts the quality of unseen translations. As quality labels, previous work uses either real valued scores estimated by humans, which require for a given QE system to address the task as a regression problem, or likert scale discrete values, which allow for the task to be addressed as either a regression or a classification problem.

Sentence-level QE has been covered by shared tasks organised by WMT since 2012, with subsequent years covering also word and document-level tasks. Recent advances in Distributional Semantics have been showing promising results in the context of QE strategies for different prediction levels. An example of that are modern word embedding architectures, such as the CBOW and Skip-Gram models introduced by (Mikolov et al., 2013b), which have been used as features in some of the best ranking systems in the sentence and word-level QE shared tasks of WMT15 (Bojar et al., 2015). Word embeddings are not only versatile, but also cheap to produce, making for both reliable and cost-effective QE solutions.

Neural Networks have also been successfully employed in QE. The FBK-UPV-UEdin (Bojar et al., 2014) and HDCL (Bojar et al., 2015) systems are good examples of that. They achieved 1st and 2nd places in the word-level QE tasks of WMT14 and WMT15, respectively, outperforming strategies that resort to much more resource-heavy features. Another successful example are neural Language Models for sentence-level QE (Shah et al., 2015).

We were not able to find, however, any examples of sentence-level QE systems that combine word embedding models and Neural Networks. In this paper, we present our efforts in doing so. We introduce SimpleNets: the resource-light and language agnostic sentence-level QE systems submitted to WMT16 that exploit the principle of compositionality for QE. In the Sections that follow, we describe the sentence-level QE task of WMT16, introduce the approach used by the SimpleNets systems, and present the results obtained.

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 812–818,
Berlin, Germany, August 11-12, 2016.

2 Task, Datasets and Evaluation

SimpleNets are two systems submitted to the sentence-level QE task of WMT16. In this task, participants were challenged to predict real-valued quality scores in 0,100 of sentences translated from English into German. The translations were produced by an in-house phrase-based Statistical Machine Translation system, and were then post-edited by professional translators. The real-valued quality scores are HTER (Snover et al., 2006) values that represent the post-editing effort spent on each given translation.

The task organisers provided three datasets:

- **Training:** Contains 12,000 translation instances accompanied by their respective post-edits and HTER values.
- **Development:** Contains 1,000 translation instances accompanied by their respective post-edits and HTER values.
- **Test:** Contains 2,000 translation instances only, without their respective post-edits or HTER values.

Each instance is composed by the original sentence in English along with its translation in German. HTER scores were capped to 100. The organisers also provided 17 baseline feature values extracted using QuEst++ (Specia et al., 2015) for each dataset.

3 The SimpleNets Approach

SimpleNets aim to provide a resource-light and language agnostic approach for sentence-level QE. Our main goal in conceiving SimpleNets was to create a reliable enough solution that could be cheaply and easily adapted to other language pairs, moving away from the use of extensive feature engineering.

The SimpleNets approach was first introduced by Paetzold and Specia (2016a) as a solution to the shared task on Quality Assessment for Text Simplification of QATS 2016[1], in which participants were asked to create systems that predict discrete quality labels for a set of automatically produced text simplifications. Labels could take three values: "Good", "Ok" and "Bad". Text Simplification differs from Machine Translation in the sense that instead of attempting to transform a text written in a source language to an equivalent text written in a target language, it attempts to transform a text in a way that it becomes more easily readable and/or understandable by a certain target audience, while still retaining the text's grammaticality and meaning.

For the Quality Assessment for Text Simplification task of QATS 2016, SimpleNets used the approach illustrated in Figure 1. For training, it performed the following five steps:

1. **Decomposition:** Given a simplification and maximum n-gram size M, it obtains the n-grams with size $1 \leq n \leq M$ of both original and simplified sentences.
2. **Union:** It then creates a pool of n-grams by simply obtaining the union of n-grams from the original and simplified sentences.
3. **Attribution:** Exploiting an interpretation of the principle of compositionality, which states that the quality of a simplification can be determined by the quality of its n-grams, it assigns the quality label of the simplification instance itself to each and every n-gram in the pool.
4. **Structuring:** Using a trained word embeddings model, it transforms each n-gram into a training instance described by a matrix $M \times N$, where M is the previously mentioned maximum n-gram size, and N the size of the word embeddings used. Each of the M rows in matrix $M \times N$ represent a given word in the n-gram, and each of the N columns, its embedding values. If an n-gram is smaller than N, the matrix is padded with embedding values composed strictly of zeroes.
5. **Learning:** Training instances are then fed into a deep Long Short-Term Memory (LSTM) Recurrent Neural Network in mini-batches so that a quality prediction model can be learned.

Notice that this process yields a model that predicts the quality of individual n-grams rather than the quality of a simplification in their entirety, which is not what was required for the task. To address this, each simplification in the test set is first processed through Decomposition, Union and

[1] http://qats2016.github.io

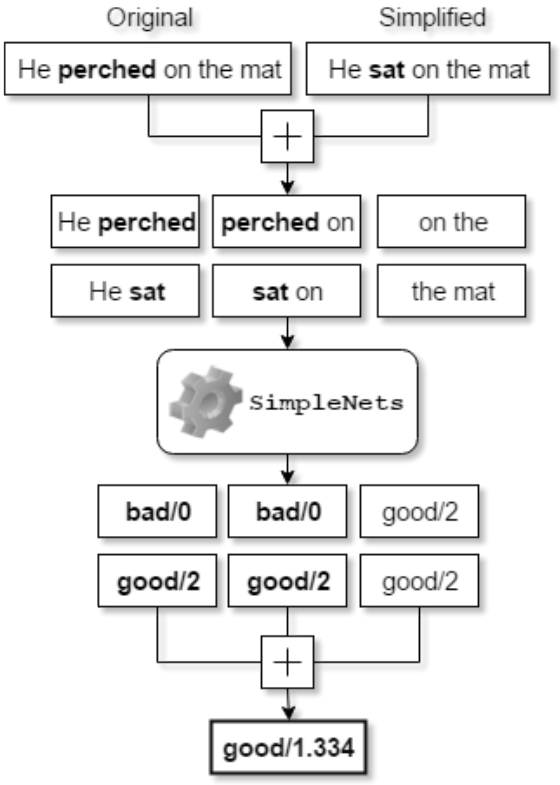

Figure 1: SimpleNets for Text Simplification

Structuring, but the complete quality prediction process has two complementary steps:

1. **Prediction:** The trained model is used to predict the quality of each n-gram of the simplification in question.

2. **Merging:** The quality of all n-grams in the simplification are merged using a certain policy, such as averaging.

After merging, a quality estimate for the simplification as a whole is produced. Although this approach has certain limitations, it addresses a very important challenge in using Recurrent Neural Networks for Text Simplification Quality Assessment: the small amount of training data available. With only 505 simplifications for training, it becomes very unlikely that a Recurrent Neural Network would be able to reliably learn a quality prediction model if it was presented with sentences in their entirety, such as how it has been done in Neural Translation and Text Generation (Schmidhuber, 2015). By splitting the sentences in the simplification in n-grams, the number of training instances available grows considerably, allowing for a better informed learning step. Additionally, the length of the sequences used in the Recurrent Neural Network becomes shorter, which can help the network to generalise the knowledge available in the training set.

The results of the Quality Assessment for Text Simplification of QATS 2016 serve as evidence of the potential of this approach: SimpleNets ranked 1st in predicting the overall quality of simplifications. Nonetheless, the inherent differences between Machine Translation and Text Simplification make it impossible for the strategy described above to be directly applied to sentence-level QE without any adaptation. In the next Section, we describe how we adapt the SimpleNets approach for sentence-level QE.

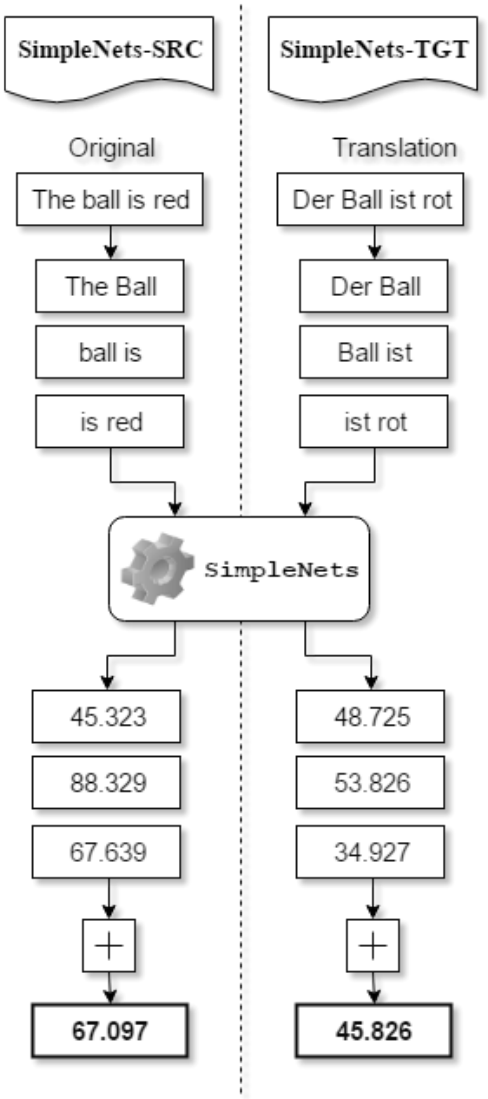

Figure 2: SimpleNets for Machine Translation

4 SimpleNets for Machine Translation

In order to use the SimpleNets strategy for sentence-level QE, we must address the biggest difference between Machine Translation and Text Simplification: while Text Simplification encompasses transformations within the constraints of a single language, Machine Translation has to handle two languages, which often have distinct vocabularies, grammar, etc. This difference prevents the application of the Union step from the process described in Section 3, since source (original) and target (translated) sentences are not in the same language, and hence cannot share the same word embeddings model during Structuring.

Another challenge in adapting SimpleNets for Machine Translation lies in the often found disparity in quality between the source sentence and its translation. Inspecting the datasets provided by the WMT16 organisers, we found that, unlike the source sentences, the majority of translations contain at least one noticeable error with respect to either grammar or coherence. This means that even if we used techniques such as the one employed by `bivec` (Luong et al., 2015), which allows for the training of bilingual word embeddings, the contrast between the quality of source and target sentences could confuse the SimpleNets approach, and hence compromise its capability of learning a reliable quality prediction model.

To overcome these challenges, we explore the hypothesis that SimpleNets can learn a better model for sentence-level QE by looking strictly at one of the sides of translations, rather than by trying to somehow combine the information from both the source and translated sentences. We train two variants of SimpleNets:

- **SimpleNets-TGT:** Explores the idea that the quality of a translation can be reliably determined based solely on the characteristics of the machine translated sentence itself, without the need to assess its relationship with the original sentence. This variant of SimpleNets aims to learn a model that is capable of quantifying the differences in quality of translated sentences.

- **SimpleNets-SRC:** Explores the idea that a translation's quality can be determined based solely on the original sentence itself, without any need to assess the intricacies of its translated version. This variant assumes that, by focusing on the original sentences and the quality scores of their translations, SimpleNets can learn how to quantify just how likely the MT system in question will be of making a mistake while attempting to translate an unseen sentence. This is in line with work on QE that explores source features to measure the complexity of the source sentence (Specia et al., 2010).

Finally, we must also address the fact that, while the quality scores provided for the QATS 2016 shared task are discrete labels, the scores for the WMT16 task are real-valued. We solve this problem by simply replacing the multiple softmax activation nodes used in the QATS 2016 SimpleNets with a single dense node, and also by replacing the cross-entropy loss function with Mean Average Error.

The workflow followed by SimpleNets-TGT and SimpleNets-SRC is illustrated in Figure 2. In the Section that follow, we describe our experiments with these approaches.

5 Experimental Setup

To assess the efficacy of our SimpleNets, we train them over the training set provided by the organizers, which contain 12,000 instances. In order to select the architecture to be used by the LSTM networks of our SimpleNets, we resort to the technique used in (Paetzold and Specia, 2016a), in which each aspect of a Neural Network is determined through parameter optimisation over the development set. The optimisation metric used is Pearson correlation, since it is the main evaluation metric adopted by the WMT16 task. The aspects of the architecture considered and the values tested for each one of them are:

1. Number of hidden layers: 1 to 5 in steps of 1.
2. Hidden layer size: 100 to 500 in steps of 100.
3. Embeddings model: CBOW or Skip-Gram.

Even though SimpleNets-TGT and SimpleNets-SRC were optimised individually, the resulting architectures of the two approaches are surprisingly the same: three hidden layers with 200 nodes each, with CBOW embeddings.

The word embedding models used were trained with `word2vec` (Mikolov et al., 2013a). We use 300 word vector dimensions and train the

System	r	MAE	RMSE
YSDA/SNTX+BLEU+SVM	0.525	12.30	16.41
POSTECH/SENT-RNN-QV2	0.460	13.58	18.60
SHEF/SVM-NN-both-emb-QuEst	0.451	12.88	17.03
POSTECH/SENT-RNN-QV3	0.447	13.52	18.38
SHEF/SVM-NN-both-emb	0.430	12.97	17.33
UGENT/SVM2	0.412	19.57	24.11
UFAL/MULTIVEC	0.377	13.60	17.64
RTM/RTM-FS-SVR	0.376	13.46	17.81
UU/UU-SVM	0.370	13.43	18.15
UGENT/SVM1	0.363	20.01	24.63
RTM/RTM-SVR	0.358	13.59	18.06
BASELINE	0.351	13.53	18.39
SHEF/SimpleNets-SRC	0.320	13.92	18.23
SHEF/SimpleNets-TGT	0.283	14.35	18.22

Table 1: Sentence-level QE scores of systems submitted to the WMT16 task

models over a corpus of around 7 billion words comprised by SubIMDB (Paetzold and Specia, 2016b), UMBC webbase[2], News Crawl[3], SUBTLEX (Brysbaert and New, 2009), Wikipedia and Simple Wikipedia (Kauchak, 2013).

For evaluation we use the task's official metrics, which are Pearson correlation (r), Mean Average Error and Root Mean Squared Error. We compare our SimpleNets with the baseline provided by the task organisers, as well as all other systems submitted. The baseline uses SVM regression with an RBF kernel and grid search for parameter optimisation.

6 Results

The task results illustrated in Table 1 reveal that SimpleNets are not as effective in sentence-level QE as they were for Text Simplification Quality Assessment. Although they outperform a few systems in terms of MAE and RMSE, when it comes to Pearson correlation, SimpleNets-SRC and SimpleNets-TGT feature at the bottom of the ranking.

What is even more surprising, however, is the difference between the performance of our SimpleNets systems. Intuitively, one would think that the n-grams of the translated sentence itself would be a more reliable indicator of a translation's quality, given that it only becomes possible for one to assess the grammaticality and meaning errors in a translation after inspecting the translated sentence itself. Interestingly, the performance scores suggest that the model employed by SimpleNets is more proficient in learning how difficult it will be for the source sentence to be translated.

The difference in performance between the SimpleNets variants became much more clear once we inspected the individual n-gram quality predictions made by them. Tables 2 and 3 show the n-grams in the development set with the highest and lowest HTER scores, as predicted by SimpleNets-TGT and SimpleNets-SRC, respectively, along the average gold HTER of the sentences in the development set which contain them. It can be noticed that the correlation between the highest scoring n-grams of SimpleNets-SRC and their average gold HTER seem to be much more pronounced than the one observed for the highest scoring n-grams of SimpleNets-TGT. The same phenomenon can be observed between the lowest scoring n-grams of the SimpleNets variants.

The Pearson correlation scores between predicted and average gold n-gram scores provide further insight on the limitations of SimpleNets in the context of sentence-level QE. While SimpleNets-TGT achieves a correlation score of 0.127, SimpleNets-SRC achieves a score of 0.151. Although SimpleNets-SRC does obtain a slightly higher Pearson score, both of them are low in comparison to other approaches, which ultimately suggests either that n-grams alone do not provide with enough information on the quality of a translation in order for a reliable Quality Estimation model to be learned, or that our method of assigning

[2]http://ebiquity.umbc.edu/resource/html/id/351

[3]http://www.statmt.org/wmt11/translation-task.html

Lowest			Highest		
N-gram	Pred.	Gold	N-gram	Pred.	Gold
das Dreieck ,	3.444	38.462	Zeile (ˆ	89.901	18.519
das Dreieck in	3.463	32.000	Vorteil dieser Methode	87.957	22.857
das Dreieck neben	3.519	11.111	Paket ist .	84.914	10.526
Dreieck , um	3.563	38.462	Lineares Licht verringert	84.042	29.412
ein Dreieck mit	3.648	76.923	einzelne Volltonfarben trennen	82.540	36.000

Table 2: N-grams with highest and lowest HTER scores, as predicted by SimpleNets-TGT

Lowest			Highest		
N-gram	Pred.	Gold	N-gram	Pred.	Gold
Backspace (Windows	2.539	10.000	gloss contour .	63.432	33.333
press Enter (	2.937	19.149	whale or white	63.432	46.154
or Option-click (	3.127	6.897	breakpoints , evaluating	63.092	57.576
Alt-click (Windows	3.128	6.897	halftone dot .	63.009	35.294
Command-D (Mac	3.397	22.857	lens focusing on	62.898	71.429

Table 3: N-grams with highest and lowest HTER scores, as predicted by SimpleNets-SRC

the translation's quality score to all n-grams during training prevents our models from learning to effectively differentiate between good and bad n-grams.

7 Final Remarks

In this paper we have described the SimpleNets systems for the sentence-level QE task of WMT16. SimpleNets aims to offer a resource-light solution to the task by exploiting Recurrent Neural Networks, word embedding models, and the principle of compositionality.

Two SimpleNets variants were described, SimpleNets-SRC and SimpleNets-TGT, which attempt to predict the quality of a translation based solely on the quality of the n-grams present in its source or target (translated) sides, respectively.

Although interesting and efficient, the SimpleNets systems have been shown not to perform well for the task at hand, featuring at the bottom of the task's final ranking. Nonetheless, our experiments have still provided with valuable insight on the impact of the source segment of a translation on the quality of its translation.

References

Nguyen Bach, Fei Huang, and Yaser Al-Onaizan. 2011. Goodness: a method for measuring MT confidence. In *Proceedings of the 49th ACL*, pages 211–219, Portland.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the 9th WMT*, pages 12–58. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the 10th WMT*, pages 1–46. Association for Computational Linguistics, September.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41:977–990.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st ACL*, pages 1537–1546.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 2015 NAACL Workshop on Vector Space Modeling for NLP*, pages 151–159, Denver, United States.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositional-

ity. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Gustavo Henrique Paetzold and Lucia Specia. 2016a. Simplenets: Evaluating simplifiers with resource-light neural networks. In *Proceedings of the 1st QATS*, pages 42–46.

Gustavo Henrique Paetzold and Lucia Specia. 2016b. Unsupervised lexical simplification for non-native speakers. In *Proceedings of The 30th AAAI*.

Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Kashif Shah and Lucia Specia. 2014. Quality estimation for translation selection. In *Proceedings of the 17th EAMT*.

Kashif Shah, Raymond W.M. Ng, Fethi Bougares, and Lucia Specia. 2015. Investigating continuous space language models for machine translation quality estimation. In *Proceedings of the 2015 EMNLP*, pages 1073–1078, Lisboa, Portugal.

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 the 2006 AMTA*, pages 223–231.

Radu Soricut and Abdessamad Echihabi. 2010. Trustrank: Inducing trust in automatic translations via ranking. In *Proceedings of the 48th ACL*, pages 612–621.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.

Lucia Specia, G Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with quest++. In *Proceedings of the 53rd ACL*, pages 115–120.

Lucia Specia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *Proceedings of the 15th EAMT*, pages 73–80.