Transparent text quality assessment with convolutional neural networks*

Robert Östling and **Gintare Grigonyte**
Department of Linguistics, Stockholm University
robert,gintare@ling.su.se

Abstract

We present a very simple model for text quality assessment based on a deep convolutional neural network, where the only supervision required is one corpus of user-generated text of varying quality, and one contrasting text corpus of consistently high quality. Our model is able to provide local quality assessments in different parts of a text, which allows visual feedback about where potentially problematic parts of the text are located, as well as a way to evaluate which textual features are captured by our model. We evaluate our method on two corpora: a large corpus of manually graded student essays and a longitudinal corpus of language learner written production, and find that the text quality metric learned by our model is a fairly strong predictor of both essay grade and learner proficiency level.

1 Introduction and related work

What makes a text good? A confluence of diverse qualities: coherent narrative, correct grammar, absence of spelling mistakes, a rich vocabulary and set of idioms. Some of these are simple to detect automatically, while others seem to require a deep understanding of the text.

Early attempts to measure text quality were pioneered by approaching it as an aggregate of distinct text features that were easy to specify manually, such as type/token ratio, average length of sentences or words, and so on. More recently, machine learning techniques have been applied that can learn such features from data.

Our primary goals in this work are to investigate how well a model for textual quality can be trained without any labeled data, and to see whether the quality model agrees with human essay graders or is able to predict second language learner proficiency.

1.1 Automated text assessment

Recent work on automated assessment mainly covers English learners' written text and it aims at assigning grades based on textual features that try to balance performance errors and language competency. Most of the work in this area falls into a category of a supervised text classification (Attali and Burstein, 2006; Landauer, 2003; Rudner and Liang, 2002; Yannakoudakis et al., 2011). Of particular interest are methods that, like ours, are based on neural networks and require little or no manual feature engineering.

1.2 Neural network approaches

Alikaniotis et al. (2016) present a model for essay scoring based on recurrent neural networks at the word level. This is trained by supervision from a graded essay corpus, and allows basic visualization of the contribution of individual words on the overall grade through error gradients. Dong and Zhang (2016) similarly train a hierarchical neural network that encodes word sequences to sentence representations, and sentence representations to essay representations, in both cases through convolution and pooling layers. The same type of approach is taken by Taghipour and Ng (2016), who however explore a wider range of models.

Cummins et al. (2016) exploit external resources through multi-task learning for automated essay scoring. This is also one of our primary motivations, but our methods are quite different.

Our method is based on deep convolutional neural networks with residual connections, which

* The source code for our system is available at https://github.com/robertostling/bea12-textquality

Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pages 282–286
Copenhagen, Denmark, September 8, 2017.

have recently gained popularity in natural language processing (Östling, 2016; Bjerva et al., 2016; Johnson and Zhang, 2016; Conneau et al., 2017).

2 Model

Since one of our primary concerns is transparency, we choose a fixed-width convolutional neural network so that it is easy to infer how each part of the text contributes to the model's estimate. In short, the whole text is passed through a one-dimensional convolutional network with residual connections, followed by a global mean pooling operation and a single fully connected layer which produces a scalar prediction of text quality. We now proceed to describe this in more detail.

Assume that the input text is a sequence of symbols (in our case characters) $s_1, s_2, \ldots, s_N$. Each symbol is represented by a row in an embedding matrix W_e of size $V \times d$, where V is the vocabulary size and d is the dimensionality of the embeddings. For convenience, we denote the embedding vector of s_i by w_i.

The sequence $w_1, w_2, \ldots, w_N$ is passed through a number of blocks with one-dimensional convolutions and residual connections (He et al., 2016). For simplicity, we let the sequence length and number of filters remain constant throughout the network (in our experiments, 512). For the first block, we use kernels of size 3, 5, 7 and 9 in order to capture character n-grams of varying size. The outputs of these are concatenated for each position in the text, similar to the encoder used by Lee et al. (2016) for character-level machine translation. This is followed by a number of blocks with only size-3 kernels. All our models use 10 blocks in total, each containing two convolutions with batch normalization layers (Ioffe and Szegedy, 2015) and rectifier non-linearities following each convolution. Let the vector x_i^l be the d-dimensional output after layer l at text position i. The final quality score of a text is computed as $q(s_{1\ldots N}) = W_o \cdot \frac{1}{N} \sum_i^N x_i^L$, that is, the dot product of the output weight vector W_o and the mean value of the outputs at the final residual layer L. In our experiments, $L = 10$.

This structure implies that the model's score for a text is the mean score over each symbol, which means that the score $q(s_{i\ldots j})$ can be computed for any subsequence $s_{i\ldots j}$ of a text without depending on the length of the sequence. This allows visualizing the low- and high-scoring sections of a text by coloring it according to the local scores.

2.1 Training

We base our model training on pairwise comparison between text snippets from different corpora or authors. We use a pseudo-probabilistic framework, where the probability of text a being better than b is defined as $P(a > b) = \sigma(q(a) - q(b))$, where $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the logistic function and $q(\cdot)$ is the quality score from our network, as detailed above. We should point out here that "better" is used from the perspective of formal written Swedish, and that "poor" text could either be informal, or due to lack of competence. During training we use cross-entropy loss, with the following axioms:

1. $P(a > b) = 0$ if a is user-generated text (**Blogs**) and b is professional prose (**News** or **SUC**).
2. $P(a > b) = 0.5$ if both a and b are professional prose.
3. $P(a > a') = 0.5$ if $\langle a, a' \rangle$ is a pair of blog texts from the same author.
4. $P(a > b) = \sigma(q(a') - q(b'))$ if $\langle a, a' \rangle$ and $\langle b, b' \rangle$ are pairs of blog texts, such that $\langle a, a' \rangle$ is from one author and $\langle b, b' \rangle$ is from another.

In plain English, these could be summarized as three general assumptions: *all authors (professional or not) are consistent*, *professional authors are better than blog authors*, and *all professional authors are equal*. Furthermore, the motivation behind point 4 is that blog authors are *not* equal, so that we can exploit the variation among them.

We initialize all model parameters, including the embeddings, randomly (orthogonal matrices for recurrent connections, Gaussian distributions for all other parameters). Due to time constraints, we did not perform hyperparameter tuning and used conservative values that worked well for similar tasks in the literature.

We train our model with stochastic gradient descent using Adam (Kingma and Ba, 2014) for learning rate adaptation. The system is implemented with Chainer (Tokui et al., 2015). In our experiments we use mini-batches of size 16, and choose an equal number of examples based for each axiom used. All text samples during training are 512 characters long. We train models for

two configurations: one using all axioms, and one only using 1+2. For the examples using axiom 4, we use a two-step procedure where the model is first use to compute $\sigma(q(a') - q(b'))$, which is then used as ground truth for those examples. We also take care to sample examples for axiom 2 from different corpora, to ensure that the model sees as different examples as possible of the same quality, avoiding that domain-specific vocabulary is mistaken for quality predictors.

2.2 Data

For model training, we use three different raw text corpora (**Blogs**, **News** and **SUC**) described below. For evaluation, we use a corpus of student essays with human-assigned grades (**Essays**), and a corpus of learner Swedish (**ASU**).

Blogs 6 billion tokens of Swedish blog posts, crawled from the web. The available metadata indicates which blog each post is sourced from, so that we can group the posts by author (assuming one author per blog).

News 100 million tokens of crawled Swedish news articles and opinion pieces, crawled from the web.

SUC 7 million tokens of published text of various genres from the Stockholm-Umeå Corpus (Källgren, 2006). This includes news, novels and academic texts.

Essays A corpus of Swedish high-school essays described in (Östling et al., 2013), containing 1,702 essays with a total of 1,1 million tokens. The data is from Swedish high school students (around age 17) with native or near-native command of Swedish. Each essay has two grades assigned by two independent human graders. While these generally have low agreement (Cohen's $\kappa = 0.399$), this is mainly due to a systematic bias by teachers assigning higher grades to their own students. We use the mean of the two grades in our analysis. Since the grading criteria mainly focus on the quality of the written language, we use this grade as a proxy for text quality.

ASU The ASU corpus (Hammarberg, 2010) is a longitudinal corpus of university-level learners of Swedish, containing two texts per session, from 11 sessions with 10 students. The progress of students is tracked from the absolute beginner stage to a level acceptable for Swedish university studies, after one or two years. The total size is about 50,000 tokens.

3 Experimental Setup and Results

We train two models, as described in Section 2.1: one using only the professional-amateur distinction (axioms 1+2) and one also using the variation in the blog corpus (axioms 1+2+3+4). The former turns out to be very poor at estimating text quality, and is only briefly discussed in Section 3.2. For the rest of this section, the 1+2+3+4 model is used throughout.

3.1 Qualitative evaluation

To illustrate the transparency of the model, Table 1 contains example sentences sampled from two text corpora (**Blogs** and **News**). In general we can see that the news examples are ranked higher than the blog examples, which is to be expected since the model was trained in part to distinguish between these corpora. The only exception is the second news sentence, whose score the visualization indicates is pulled down by the first word, 'domen' (*the sentence*). This turns out to be a homograph of 'dom', a spoken-language form of the third person plural pronoun, which is generally avoided in written Swedish and a strong indicator of either an informal style or poor command of Swedish (since the written language makes a case distinction which does not exist in the modern spoken language). Other low-scoring features include smileys, frequent use of ellipsis, and informal spellings such as 'oxå' for 'också' (*also*). Some of these are typical for informal Internet text, and would easily be avoided in e.g. a high-stakes essay setting. However, rather than low scores stemming from occasional features of poor or informal writing, it seems that the consistent lack of a richer vocabulary is a more important factor.

3.2 Native language essay grades

We compute the scores for each of the 1,702 essays in the **Essays** dataset. Since the essays were produced during a fixed-time test situation, length is a strong predictor of grade ($R^2 = 0.308$ for the 4th root of essay length in characters, $L^{0.25}$; we report adjusted R^2 from multiple linear regression). Controlling for length, the 1+2 model is not a significant predictor of grade. The 1+2+3+4

Table 1: Mean scores (left) and color-coded partial scores (right) for a sample of sentences from the blog corpus (top) and news corpus (bottom). Red encodes low scores, blue encodes high scores. Faded colors are used for scores near zero.

	Blogs
1.475	Resten av veckan blir det jag som VAB:ar. Men det tycks inte bli så tråkigt som det låter...
0.530	Och på torsdag ska vi hem till Neos dagiskompis som oxå åkt på skiten, yeey!
-0.256	Någon mer som vill leka med oss och kanske bli smittad? Bara att hojta! :D
0.143	Spännande att få äta med sked och känna liiiite motstånd i munnen för en gång skull, haha! :D
	News
2.310	Rättegången mot Geert Wilders direktsänds i holländsk tv. Det hör inte till vanligheterna.
0.613	Domen väntas i början av november.
2.428	I Storbritannien finns fem miljoner katoliker, vilket motsvarar en tolftedel av befolkningen.
2.611	Allt annat skulle betyda att det nyvunna förtroendet för Lettland går förlorat.

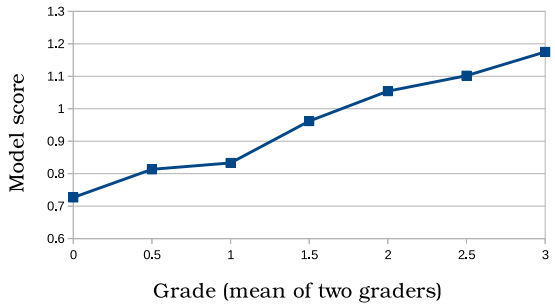

Figure 1: Relation between human-assigned grades and scores from our model.

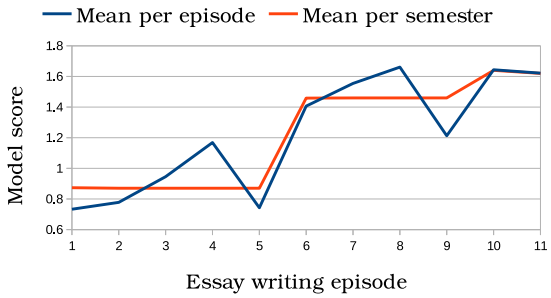

Figure 2: The progress of Swedish learner essay scores' during 11 writing episodes. Both curves display the same data, but averaged over writing episodes or semesters (i.e. down-sampled to smooth the curve), respectively.

model is a moderately strong predictor of grade ($R^2 = 0.127$ on its own, $R^2 = 0.355$ together with $L^{0.25}$).

The relation between essay grade and model score is illustrated in Figure 1, where for each of the seven possible grade means (0.0–3.0 in half-point intervals) the mean score of all essays with that grade is shown.

3.3 Second language learner progress

We use the **ASU** corpus (Hammarberg, 2010) to investigate whether our model can estimate the progress made by second-language learners during their early stages of acquiring Swedish as a second language.

Figure 2 shows how our model's score changes over the 11 sessions that the participants took part in. We compute the scores by pooling the essays from each session (20 essays, 2 each for 10 students). There is a clear increasing trend.

4 Conclusions

We have presented a model based on deep convolutional neural networks, which is able to estimate text quality at both the local and global scale, allowing easy visualization of weak or strong points of the text. Our method is using only unlabeled text corpora as training data, but its predictions align well with human-assigned grades for native-language essays and the time progression for second language learners. We expect this to be a useful component in systems for automated essay scoring and feedback.

Acknowledgments

We thank the reviewers for their constructive comments. Computing resources were provided by the Finnish IT Center for Science (CSC).

References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 715–725. http://www.aclweb.org/anthology/P16-1068.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment* 4(3).

Johannes Bjerva, Barbara Plank, and Johan Bos. 2016. Semantic tagging with deep residual networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 3531–3541. http://aclweb.org/anthology/C16-1333.

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2017. Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, pages 1107–1116. http://www.aclweb.org/anthology/E17-1104.

Ronan Cummins, Meng Zhang, and Ted Briscoe. 2016. Constrained multi-task learning for automated essay scoring. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 789–799. http://www.aclweb.org/anthology/P16-1075.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring – an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1072–1077. https://aclweb.org/anthology/D16-1115.

Björn Hammarberg. 2010. Introduction to the asu corpus: A longitudinal oral and written text corpus of adult learner swedish with a corresponding part from native swedes. white paper. version 2010-11-16. .

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. *CoRR* abs/1603.05027. http://arxiv.org/abs/1603.05027.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In David Blei and Francis Bach, editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. JMLR Workshop and Conference Proceedings, pages 448–456. http://jmlr.org/proceedings/papers/v37/ioffe15.pdf.

Rie Johnson and Tong Zhang. 2016. Convolutional neural networks for text categorization: Shallow word-level vs. deep character-level. *CoRR* abs/1609.00718. http://arxiv.org/abs/1609.00718.

Gunnel Källgren. 2006. *Manual of the Stockholm Umeå Corpus version 2.0*. Department of Linguistics, Stockholm University. Sofia Gustafson-Capková and Britt Hartmann (eds.).

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980. http://arxiv.org/abs/1412.6980.

Thomas K Landauer. 2003. Automatic essay assessment. *Assessment in education: Principles, policy & practice* 10(3):295–308.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2016. Fully character-level neural machine translation without explicit segmentation. *CoRR* abs/1610.03017. http://arxiv.org/abs/1610.03017.

Robert Östling. 2016. Morphological reinflection with convolutional neural networks. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. http://aclweb.org/anthology/W/W16/W16-2003.pdf.

Robert Östling, André Smolentzov, Björn Tyrefors Hinnerich, and Erik Höglin. 2013. Automated essay scoring for Swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Atlanta, Georgia, pages 42–47. http://www.aclweb.org/anthology/W13-1705.

Lawrence M Rudner and Tahung Liang. 2002. Automated essay scoring using bayes' theorem. *The Journal of Technology, Learning and Assessment* 1(2).

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1882–1891. https://aclweb.org/anthology/D16-1193.

Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. 2015. Chainer: a next-generation open source framework for deep learning. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Human Language Technologies-Volume 1*. ACL, pages 180–189.