Automated Evaluation of Scientific Writing: AESW Shared Task Proposal

Vidas Daudaravičius
VTeX
Mokslininku st. 2a
Vilnius, Lithuania
vidas.daudaravicius@vtex.lt

Abstract

The goal of the Automated Evaluation of Scientific Writing (AESW) Shared Task is to analyze the linguistic characteristics of scientific writing to promote the development of automated writing evaluation tools that can assist authors in writing scientific papers. The proposed task is to predict whether a given sentence requires editing to ensure its "fit" with the scientific writing genre. We describe the proposed task, training, development, and test data sets, and evaluation metrics.

Quality means doing it right when no one is looking.
– Henry Ford

1 Introduction

De facto, English is the main language for writing and publishing scientific papers. In reality, the mother-tongue of many scientists is not English. Writing a scientific paper is likely to require more effort for researchers who are nonnative English speakers compared to native speakers. The lack of authoring support tools available to nonnative speakers for writing scientific papers in English is a formidable barrier nonnative English-speaking authors who are trying to publish, and this is becoming visible in academic community. Many papers, after acceptance to journals, require improvement in overall writing quality which may be addressed by publishers. However, this is not the case with most conference proceedings.

The vast number of scientific papers being authored by nonnative English speakers creates a large demand for effective computer-based writing tools to help writers compose scientific articles. Several shared tasks have been organized (Dale and Kilgarriff, 2011; Dale et al., 2012; Ng et al., 2013; Ng et al., 2014) which constituted a major step toward evaluating the feasibility of building novel grammar error correction technologies. English language learner (ELL) corpora were made available for research purposes (Dahlmeier et al., 2013; Yannakoudakis et al., 2011). An extensive overview of the feasibility of automated grammatical error detection for language learners was conducted by Leacock et al. (2010). While these achievements are critical for language learners, we also need to develop tools that support genre-specific writing features. The shared task proposed here focuses on the genre of scientific writing.

Above and beyond correct use of English conventions, the genre of scientific writing is characterized by features, including, but not limited to declarative voice, and appropriate academic and discipline-specific terminology. There are many issues for writers that are not necessarily related to grammar issues such as, vocabulary usage, and correct word and phrase order among other issues. In addition, many ELL writers have a different way of thinking and reasoning in their native language which may be reflected in their writing. For instance, it is likely that ELLs and native English (EN) writers would write the same text in different ways:

1. ELL *"Completely different role of elastic interaction occurs due to local variations in the strain field..."*

   EN *"Elastic interaction takes on a completely*

Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015, pages 56–63,
Denver, Colorado, June 4, 2015.

different role with local variations in the strain field..."

2. ELL *"The method is straightforward and concise, and its applications is promising."*

   EN *"The method is straightforward and concise, and it holds promise for many applications."*

The difference in the readability and the fluency of texts due to grammatical errors is apparent.

The task of automated writing evaluation applied to scientific writing is critical, but it is not well studied because no data for research have been available until recently when the dataset of language edits of scientific texts was published (Daudaravicius, 2014).

On the other hand, some scientists propose to use Scientific Globish versus scientific English (Tychinin and Kamnev, 2013). The term 'Globish' denotes the international auxiliary language proposed by Jean-Paul Nerrière, which relies on a vocabulary of 1500 English words and a subset of standard English grammar[1]. The proposed adoption of '*scientific Globish*' as a simplified language standard may appeal to authors who have difficulty with English proficiency. However, *Globish* might lead to further deterioration of the quality of English-language scientific writing, and, in general, it cannot be a reasonable direction. Therefore, we propose the *automated evaluation of scientific writing* shared task.

2 Language Quality in Scientific Discourse

In this section, we define the concept of *language quality* and provide examples of previous work that has evaluated scientific writing.

2.1 Definition

While writers may have proficiency in English, they may still struggle to be effective writers in the genre of scientific writing. The concept of '*quality*' in scientific discourse is ill-defined. For instance, a student in a seventh-grade science classroom asked a question '*Maestro, what is quality?*' during an experiment engaging students to address two questions: "*What is the quality of air in my community?*" and "*What is the quality of water in our river?*" (Moje et al., 2001). The student was asking, "*What do you mean when you talk about quality?*" As a result of this question, Maestro Tomas spent a class period working on what it meant to refer to quality, especially in science, and how scientists determined *quality*. In the most explicit discussion, Maestro Tomas told the students that *quality* differs depending on one's purpose, one's background, and one's position (e.g., as a scientist, an activist, an industrialist, a community member).

We find that the concept of *academic language* and the concept of *the language of academic writing* are different at a conceptual level. Krashen and Brown (2007) discuss the concept of academic language proficiency. They argue that academic language proficiency consists of the knowledge of academic language and specialized subject matter. The *academic language* concept can be described as a proper use of discipline-specific and academic vocabulary to express topic and discourse structure.

2.2 Previous work: Scientific Writing Evaluation

Natural language software requirements are the communication medium between users and software developers. Ormandjieva et al. (2007) addressed a problem of writing evaluation of natural language software requirements, and applied a text classification technique for automatic detection of ambiguities in natural language requirements. Sentences were classified as "ambiguous" or "unambiguous", in terms of surface understanding. Fabbrini et al. (2001) present a tool called QuARS (Quality Analyzer of Requirements Specification) for the analysis of textual software requirements. The Quality Model aims at providing a quantitative, corrective and repeatable evaluation of software requirement documents. Berrocal Rojas and Sliesarieva (2010) examine the automated detection of language issues affecting accuracy, ambiguity and verifiability in natural language software requirements. Lexical analysis, syntactic analysis, WordNet (Miller et al., 1993) and VerbNet (Schuler, 2005) were used for the automated quality evaluation. Burchardt et al. (2015) provided practical guidelines for the use of the Multidimensional Quality Metrics (MQM) framework for assessing translation quality in scientific research projects. MQM provide detailed

[1] See: http://en.wikipedia.org/wiki/Globish_(Nerriere)

The boundary problem for $V(t,x)$ is of the form

$$(\partial_t + L - r)V(t,x) = 0, \quad x > h, t < T; \tag{1}$$
$$V(t,x) = 0, \quad x \le h, t \le T; \tag{2}$$
$$V(T,x) = G(x), \quad x > h. \tag{3}$$

Boyarchenko and Levendorskiĭ (BLbook; BLAAP02) derived the generalization of the Black–Scholes equation 1 under a weak regularity condition: the process (t, X_t) in 2D satisfies the (ACP) condition (for the definition, see e.g. (Sa)). Note that the (ACP) condition is satisfied if the process X has a transition density. Equation 1 is understood in the sense of the theory of generalized functions: for any infinitely smooth function u with compact support $\operatorname{supp} u \subset (-\infty, T) \times (h, +\infty)$,

$$(V, (-\partial_t + \tilde{L} - r)u)_{L_2} = 0, \tag{4}$$

where $\tilde{L}$ is the infinitesimal generator of the dual process.

Figure 1: A short example of common academic text writing (from (Kudryavtsev and Levendorskiĭ, 2009)).

The boundary problem for _MATH_ is of the form _MATHDISP_. Boyarchenko and Levendorskii _CITE_ derived the generalization of the Black–Scholes equation (_REF_) under a weak regularity condition: the process _MATH_ in 2D satisfies the (ACP) condition (for the definition, see e.g. _CITE_). Note that the (ACP) condition is satisfied if the process _MATH_ has a transition density. Equation (_REF_) is understood in the sense of the theory of generalized functions: for any infinitely smooth function _MATH_ with compact support _MATH_, _MATHDISP_, where _MATH_ is the infinitesimal generator of the dual process.

Figure 2: The transformation of the text in Fig 1 using named entities.

insights about translation issues/errors on different levels of granularity up to the word or phrase level as input for systematic approaches to overcome translation quality barriers. MQM framework does not provide a translation quality metric, but rather provides a framework for defining task-specific translation metrics. MQM describes three typical layers of annotation in MT development:

- the phenomenological level (target errors/issues);
- the linguistic level (source or target POS, phrases, etc.);
- the explanatory level (source/system-related causes for certain errors).

A wide range of translation quality evaluation aspects show that the field is growing, and more efforts needed to solve many issues of translation quality evaluation.

3 The Language of Scientific Texts

Some elements of scientific writing that are distinct from other genres of writing, include, but are not limited to the following:

- Formal notations, e.g. $f(x) = \cos(x)$.
- Extensive mathematical expressions which can be independent sentences or a continuation of a preceding sentence, see example in Fig 1.
- Discipline-specific terminology.
- Citations.
- Section headers.
- References to other elements of a paper, which are of logical relation only. The scientific writing is highly multidimensional compared to linear daily language.
- Lists and enumerations.
- Bibliography elements.

Domain	The Number of Paragraphs	The Number of Edits
Physics	41,188	164,813
Mathematics	32,981	79,019
Engineering	14,968	43,551
Statistics	12,115	35,988
Computer Science	7,028	16,013
Astrophysics	4,278	15,594
Business and Management	3,454	8,262
Psychology	2,604	6,189
Finance	2,241	6,016
Economics	185	314
Total	**121,042**	**375,759**

Table 1: Main characteristics of the training dataset.

– Figures are also used as the continuation of sentences, though not so frequently.

– Hypertext references.

4 The Task Objectives and Definition

The objectives of the AESW Shared Task are to promote the use of NLP tools to help ELL writers the quality of their scientific writing.

In the scope of the task, the main goals are:

– to identify sentence-level features that are unique to scientific writing;

– to provide a common ground for development and comparison of sentence-level automated writing evaluation systems for scientific writing;

– to establish the state-of-the-art performance in the field.

Some interesting uses of sentence-level quality evaluations are the following:

– automated writing evaluation of submitted scientific articles;

– authoring tools in writing English scientific texts;

– filtering out sentences that need quality improvement.

The task will examine automated evaluation of scientific writing at the sentence-level by using the output of the professionally edited scientific texts, which are text extracts before and after editing (by native English speakers).

The goal of the task is to predict whether a given sentence needs for any kind of editing to improve it. The task is a binary classification task. Two cases of decisions are examined: binary decision (False or True) and probabilistic estimation (between 0 and 1).

5 Data

5.1 The Editing Process

This section describes the role of the professional language editors who completed the data editing described in Section 5.3. *Language editors* are defined as individuals who perform *proofreading* (see Smith (2003)). There are no standards that define language quality. The language editors use best practices, for instance (see Society for Editors and Proofreaders (2015)).

Language editors edited selected papers as part of publishing service. Each edited paper has two versions: text *before* and *after* editing. Language editors do their best to improve writing quality within the limited time span. In this data set, however, there was no double-annotation for quality control. We estimate that approximately 20% of the data may still contain errors, and also that there may be errors in the editors edits.

5.2 Tex2TXT

We use the open-source tool `tex2txt`[2] for the conversion from LaTeX to text, which was developed

[2]See: http://textmining.lt:8080/tex2txt.htm

```
<par pid="9" domain="Physics">
    <edits>
        <edit originalParOffset="7" editedParOffset="7" type="replaced">
            <original>ultimately</original>
            <edited>finally</edited>
        </edit>
    </edits>
    <sentence type="original" sid="9.0">Let us ultimately insist on the fact that the expression in the right hand side _MATH_
                                        is a function of _MATH_ due to the action of the shift and is therefore a different
                                        function than _MATH_. </sentence>
    <sentence type="edited" sid="9.1">Let us finally insist on the fact that the expression in the right hand side _MATH_
                                        is a function of _MATH_ due to the action of the shift and is therefore a different
                                        function than _MATH_. </sentence>
    <sentence type="nonedited" sid="9.2">Only the expectations of both expressions of Eq. (_REF_) are equal.</sentence>
</par>
```

Figure 3: Training data example of the paragraph annotation with data before language editing, after language editing, and the difference.

```
<par pid="9" domain="Physics">
    <sentence sid="9.0">Let us ultimately insist on the fact that the expression in the right hand side _MATH_ is a func-
                        tion of _MATH_ due to the action of the shift and is therefore a different function than _MATH_.
                        </sentence>
    <sentence sid="9.1">Only the expectations of both expressions of Eq. (_REF_) are equal.</sentence>
</par>
```

Figure 4: A sample from the test data.

specifically for this task. The tool is stand-alone and does not require any other LaTeX processing tools or packages. The primary goal was to extract the correct textual information.

5.3 The Data Set

The data set is the collection of text extracts from more than 4,000 published journal articles (mainly from physics and mathematics) *before* and *after* language editing. The data were edited by professional editors (per above) who were native English speakers[3]. Editing includes grammar error corrections, text cleaning, rephrasing, spelling correction, stylistics, and sentence structure corrections. Each extract is a paragraph which contains at least one edit done by language editor. All paragraphs in the dataset were randomly ordered for the source text anonymization purpose. The distribution of paragraphs and edits are presented in Table 1.

Sentences were tokenized automatically, and then both versions – texts *before* and *after* editing – automatically aligned with a modified `diff` algorithm. Each sentence is annotated as either '*original*', or '*edited*', or '*nonedited*'. *Non-edited* sentences contained no errors. The *original text* – the text before language editing – can be restored simply by deleting sentences that are annotated as 'edited'. Also, the *edited text* can be restored simply by deleting sentences that are annotated as 'original'.

The training data: The training data will be at least 121,000 paragraphs with 375,000 edits. The number of edited sentences will be at least 235,000, and the number of original sentences will be at least 234,000. There will be 335,000 sentences that were non-edited. These numbers show that 41% of all sentences were edited. See

[3]VTeX provides LaTeX-based publishing solutions and data services to the scientific community and science publishers. Publishers often request language editing services for papers accepted for publication. The data of our proposed shared task are based on selected papers published in 2006–2009 by Springer publishing company and edited at VTeX by professional language editors.

Figure 3 for an example of annotated training data.

The training data will include annotations to show differences between the 'original' and 'edited' texts. The 'edits' data are used for a quick reference to what the changes are.

The development data: An additional 5,000 paragraphs similar to test data will be provided. The development data set will be comprised of a set of articles that are independent from articles used for compiling the training and test sets. The development data will be distributionally similar to training data and test data with regard to edited and non-edited sentences, and domain.

The test data: An additional 5,000 paragraphs will be provided for testing the registered systems of the AESW Shared Task. The test data set will be comprised of a set of articles that are independent from articles used for compiling the training and development sets. Test paragraphs will retain 'original' and 'nonedited' versions only. The 'edited' sentence version will be removed. The test data anotation will be similar to training and development data. However, no data about edits and sentence class will be provided until submission of system results. See an example in Figure 4.

Shared Task participating teams will be allowed to use external data that are publicly available. Teams will not be able to use proprietary data. Use of external data should be specified in the final system report.

6 The Task and Evaluation

The task is to predict the class of a test sentence: 'original' or 'edited'. In Section 2, we saw that both Boolean and probabilistic prediction are used for various tasks. Therefore, there will be **two tracks of the task**:

Boolean Decision: The prediction of whether a test sentence is edited (TRUE), or before editing and corrections are needed (FALSE).

Probabilistic Estimation: The probability estimation of whether a test sentence is edited ($P = 0$), or before editing and corrections are needed ($P = 1$).

Participating teams will be allowed to submit up to two system results for each track. In total, a maximum of four system results will be accepted. All participating teams are encouraged to participate in both tracks.

The primary goal of the task is to predict 'original' sentences with poor writing quality. Each registered system will be evaluated with a *Detection score*, which is described below.

6.1 Detection score

The score will be an F-score of 'original' class prediction. The score will be computed for both tracks individually. For the Boolean decision track, a gold standard sentence G_i is considered detected if there is an alignment in the set that contains G_i. We calculate *Precision* (*P*) as the proportion of the sentences that were 'original' in the gold standard:

$$P_{bool} = \frac{\#\ Sentence_{detected}}{\#\ Sentence_{spurious} + \#\ Sentence_{detected}}.$$

Similarly, *Recall* (*R*) will be calculated as:

$$R_{bool} = \frac{\#\ Sentence_{detected}}{\#\ Sentence_{gold}}.$$

The *detection score* is the harmonic mean (F-score):

$$DetectionScore_{bool} = 2 \cdot \frac{P_{bool} \cdot R_{bool}}{P_{bool} + R_{bool}}.$$

For the probabilistic estimation track, the Mean squared error (MSE) will be used. A gold standard sentence G_i is assigned to 1 if it is 'original', and to 0 if it is 'nonedited'. A gold standard sentence G_i is considered detected if there is correlation in the set that contains G_i. We calculate *Precision* as the MSE of the sentences E_i that were estimated as 'original', i.e., their estimated probability is above 0.5:

$$P_{prob} = 1 - \frac{1}{n}\sum_{i=1}^{n}(E_{i,\,>0.5} - G_i)^2.$$

The higher the P_{prob} the better the system is. Similarly, we calculate *Recall* as the MSE of the sentences G_i that were 'original' in the gold standard:

$$R_{prob} = 1 - \frac{1}{n}\sum_{i=1}^{n}(E_i - G_{i,original})^2.$$

ID	Type	G_{bool}	G_{prob}	Boolean Decision Track			Probabilistic Estimation Track		
				TEAM1	TEAM2	TEAM3	TEAM1	TEAM2	TEAM3
1	original	F	1	F	T	F	0.7	0	1
2	original	F	1	F	T	F	0.8	0	1
3	nonedited	T	0	T	T	F	0.1	0	1
4	nonedited	T	0	F	T	F	0.6	0	1
5	nonedited	T	0	T	T	F	0.2	0	1
6	nonedited	T	0	T	T	F	0.4	0	1
7	original	F	1	F	T	F	0.9	0	1
8	nonedited	T	0	T	T	F	0.1	0	1
9	nonedited	T	0	T	T	F	0.4	0	1
P				**0.75**	**0**	**0.33**	**0.875**	**0**	**0.33**
R				**0.67**	**0**	**1**	**0.953**	**0**	**1**
DetectionScore				**0.71**	**0**	**0.5**	**0.912**	**0**	**0.5**

Table 2: $DetectionScore$ calculation example.

The harmonic mean $DetectionScore_{prob}$ is calculated similarly as $DetectionScore_{bool}$. The higher the $DetectionScore_{prob}$ the better the system is. An example of score calculation is shown in Table 2.

7 Report submission

The authors of participant systems are expected to submit a shared task paper describing their system. The task papers should be 4-8 pages long and contain a detailed description of the system and any further insights.

Acknowledgments

We would like to thank the Springer publishing company for permission to publish a large number of text extracts from published scientific papers. We appreciate the support and help in improving writing quality and organization of this paper from Building Educational Applications (BEA) workshop organisers. And special thanks to Joel Tetreault for the discussions and valuable suggestions. This work has been partially supported by EU Structural Funds administrated by the Lithuanian Business Support Agency (Project No. VP2-1.3-UM-02-K-04-019).

References

Allan Berrocal Rojas and Gabriela Barrantes Sliesarieva. 2010. Automated detection of language issues affecting accuracy, ambiguity and verifiability in software requirements written in natural language. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 100–108, Los Angeles, CA.

Aljoscha Burchardt, Kim Harris, Alan K. Melby, and Hans Uszkoreit, 2015. *Multidimensional Quality Metrics (MQM) Definition*. Version 0.3.0 (2015-01-20), http://www.qt21.eu/mqm-definition/definition-2015-01-20.html.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS Corpus of Learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, GA, June.

Robert Dale and Adam Kilgarriff. 2011. Helping Our Own: The HOO 2011 Pilot Shared Task. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, pages 242–249, Nancy, France.

Robert Dale, Ilya Anisimoff, and George Narroway. 2012. HOO 2012: A report on the Preposition and Determiner Error Correction Shared Task. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 54–62, Montréal, Canada.

Vidas Daudaravicius. 2014. Language editing dataset of academic texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

F. Fabbrini, M. Fusani, S. Gnesi, and G. Lami. 2001. An automatic quality evaluation for natural language requirements. In *Proceedings of the Seventh International Workshop on RE: Foundation for Software Quality (REFSQ2001)*, pages 4–5.

Stephen Krashen and Clara Lee Brown. 2007. What is academic language proficiency? *STETS Language & Communication Review*, 6(1):252–262.

Oleg Kudryavtsev and Sergei Levendorskiĭ. 2009. Fast and accurate pricing of barrier options under Lévy processes. *Finance and Stochastics*, 13(4):531–562.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*. Morgan and Claypool.

G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1993. Introduction to WordNet: An online lexical database. Technical report.

Elizabeth B. Moje, Tehani Collazo, Rosario Carrillo, and Ronald W. Marx. 2001. "Maestro, what is 'quality'?": Language, literacy, and discourse in project-based science. *Journal of Research in Science Teaching*, 38(4):469–498.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, MD, USA.

Olga Ormandjieva, Ishrar Hussain, and Leila Kosseim. 2007. Toward a text classification system for the quality assessment of software requirements written in natural language. In *Fourth International Workshop on Software Quality Assurance: In Conjunction with the 6th ESEC/FSE Joint Meeting*, SOQUA '07, pages 39–45, New York, NY, USA. ACM.

Karin Kipper Schuler. 2005. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, Philadelphia, PA, USA.

B. Smith. 2003. *Proofreading, Revising & Editing Skills Success in 20 Minutes a Day*. Learning Express Library. Learning Express.

Society for Editors and Proofreaders, 2015. *Ensuring editorial excellence: The SfEP code of practice*. http://www.sfep.org.uk/pub/bestprac/cop5.asp.

Dmitry N. Tychinin and Alexander A. Kamnev. 2013. Scientific Globish versus scientific English. *Trends in Microbiology*, 21(10):504–505.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, OR, USA.