Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Tradares: A Tool for the Automatic Evaluation of Human Translation Quality within a MOOC Environment

Miguel Betanzos, Marta R. Costa-jussà & Lluís Belanche

To cite this article: Miguel Betanzos, Marta R. Costa-jussà & Lluís Belanche (2017) Tradares: A Tool for the Automatic Evaluation of Human Translation Quality within a MOOC Environment, Applied Artificial Intelligence, 31:3, 288-297, DOI: 10.1080/08839514.2017.1317154

To link to this article: https://doi.org/10.1080/08839514.2017.1317154

APPLIED ARTIFICIAL INTELLIGENCE
2017, VOL. 31, NO. 3, 288–297
https://doi.org/10.1080/08839514.2017.1317154

Taylor & Francis
Taylor & Francis Group

Tradares: A Tool for the Automatic Evaluation of Human Translation Quality within a MOOC Environment

Miguel Betanzos[a], Marta R. Costa-jussà[b], and Lluís Belanche[a]

[a]Department de Ciències de la Computació, Universitat Politècnica de Catalunya, Barcelona, Spain; [b]Departament de Teoria del Senyal i Comunicacions, Universitat Politècnica de Catalunya, Barcelona, Spain

Abstract

In this paper, we introduce TradARES, a tool for the automatic evaluation of human translation quality developed in the context of an OpenEdx MOOC (Massive Open Online Course), setting the foundation for a tool that provides efficient and trustful feedback to students. Our further goal is to release a small corpus of Arabic-Spanish translations from the first edition of the course. The evaluation tool is based on prediction models and at the moment is able to indicate the quality of a given piece of text—in the numerical scale {1,2,3,4}—as a translation of another piece of text.

Introduction

The explosion of the Massive Open Online Courses (MOOC) phenomenon happened in 2012 with the arrival of names like Coursera, EdX or Udacity, soon followed by many other providers (Pappano 2012). It might well be considered one of the biggest innovations in education of our time. The figures are indeed big: by 2013, the number of enrolled students was in the millions, thousands of courses had been offered, and hundreds of universities were offering their courses in this format (Christensen et al. 2013). Nowadays, browsing through MOOC catalogs, we can see that very few of these courses are offered on the subject of natural language learning. This is somewhat surprising, since this field could possibly generate a strong interest among many students.

This paper describes the process followed in the development of a tool for automatic evaluation of human translation quality in the context of Arabic to Spanish translation. The idea was to take advantage of the multiple resources found in MOOCs, and so, we prepared a 3-week translation course[1,2] that was freely offered to participants through an instance of the Open EdX platform hosted in an Amazon virtual server.

The course offers participants collaborative learning: they receive evaluations and suggestions from other participants, and they analyze the mistakes and successes in the translations of other participants, following a specific

CONTACT Marta R. Costa-jussà marta.ruiz@upc.edu Departament de Teoria del Senyal i Comunicacions, Universitat Politècnica de Catalunya, Barcelona 08034, Spain.

rubric. This is a practical way to learn as well as reflect on the mechanisms of translation. The texts are real and relevant, ranked from the least to the most difficult, and divided into areas or themes. Additionally, the course contains some exercises and support materials, and there is the possibility of discussing related topics in the forum. In the end, we were able to compile a corpus using the translations obtained during the course, with the intention of building a tool able to perform automatic evaluation of the quality of new translations. This tool relies on several linguistic features extracted from the translation corpus, and evaluations provided from the participants.

During the first 2 weeks of the course, typical translation problems were addressed, with a stress on transliteration issues specific to Arabic-Spanish translation. Each of the 2 weeks consisted of four texts, exercises related to the texts, support material, and forum discussion. In the last week of the course, students were asked to input their translation of the studied texts into the system, in an anonymous manner. Then, they received marks and comments from other participants, while in turn giving marks and comments to translations submitted by other participants, in a peer-to-peer scheme, always following a specific rubric.

Translations paired with the marks provided by students are used to create the corpus that feeds the automatic evaluation tool. This tool relies on several linguistic features extracted from the corpus, along with the evaluations provided from the participants, in order to build a prediction model. The possibility of getting immediate feedback on a given translation could be a very useful resource in the fields of language teaching and translation teaching, among others.

The rest of the paper is organized as follows: "Preliminaries" section explains the context of the present study and related work. "Materials and methods" section reports details about the data collection, the peer-review process, the extracted information, and the modeling methods. "Results and discussion" section presents and discusses the results obtained. Finally, "Conclusions" section makes a summary of the conclusions of this work.

Preliminaries

MOOCs are Web-based courses that allow anyone with an internet connection to enroll, because they are open, free, and have no maximum enrollment limits. They offer all the content or references required for the course for free, and require very less involvement of an instructor from a student perspective after the course begins (Balfour 2013). In particular, the Open EdX initiative is completely open source and can be adopted by any education institution willing to do so. For instance, Catalonian universities have created UCATx,[3] a virtual space offering free online open courses about diverse subjects, but the tendency is to join bigger platforms, as EdX itself,

and offer courses through a centralized big platform instead of deploying (and managing) a dedicated platform.

Automatic evaluation of translation quality is mostly used to assess the output of machine translation systems. It is indeed very necessary to improve the work of such systems, which can produce enormous quantities of translations that could not possibly be all assessed by human experts. As expressed by (Snover et al. 2006), "*Machine translation has proven a difficult task to evaluate. Human judgments of evaluation are expensive and noisy*."

The evaluation of translations could be categorized into two main branches: one that needs reference translations to produce an assessment, and another that only uses the source text. The latter is simpler and particularly useful for the development of prediction models, which could, in a limited way, be taken as reference translations. Recent work on this topic mainly involves finding the most informative features to extract from the text, coupled with feature selection algorithms to reduce the number of required features (Shah, Cohn, and Specia 2015).

Materials and methods

As follows, this section describes the details about the data collection, which is made available for free to the community; the peer-review process; the extracted information; and the modeling methods.

Data collection

The course ran from August 17, 2015 to September 8, 2015. It had a total of eight source texts in Arabic, each around 250 words, which had to be translated into Spanish. There were over 120 participants enrolled in the course, and the peer-reviewed exercise obtained 32 translations for text one (Text1), and 23 translations for text two (Text2). The rest of the source texts were translated by only 12 participants, which we considered too small a data set to work with, and were discarded for further analysis. Each of the translations collected has three evaluations given by peers. The Arabic source texts, their different translations into Spanish, and their evaluations are available at Research Gate.[4]

Peer review

Translation quality was measured in a peer-to-peer scheme—Figure 1 shows a detailed scheme of this assessment. Participants were then instructed to give a mark between 1 and 4 to their peers' translations, in accordance to a rubric on

Amount of edition necessary for the translation to be acceptable: How many editing should be needed in order to turn it into a working translation? (it captures the meaning of the text from the source language to the target language)

1. Almost everything should be edited. It is necessary to translate the text again. This translation does not serve its purpose.
2. A significant amount of edition is needed, but it is not necessary to translate again. It is faster to edit than to start the translation again.
3. Little edition is needed. The translation we are evaluating serves its purpose. Some changes are still needed to deem it acceptable.
4. There is no need to edit. The translation serves its purpose. There is no need to make any change or maybe just some cosmetic changes.

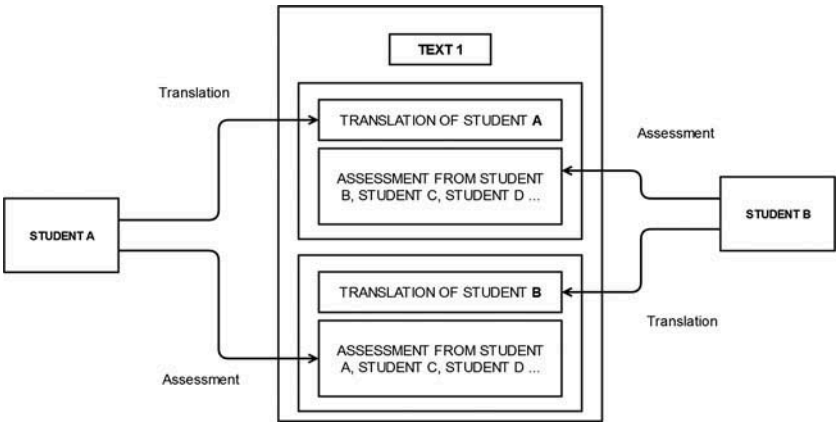

Figure 1. Peer-review assessment scheme used for evaluation.

translation edit rate, as in (Specia et al. 2009), with the help of a specially designed rubric, as follows:

As an optional part in the rubric, two additional questions were posed:

- Could you point out the reasons for your evaluation?
- What kind of errors did you find in the translation? Please specify spelling errors, grammar errors (verb tenses, gender and number concordances, etc.), translation sense errors, omissions, unfinished translations...

The idea was that each translation could also be tagged with keywords extracted from the optional part, such as "grammar," "spelling," "omissions," etc., aiming to establish a relationship between the translation features and these tags. This part did not turn out as expected because several other tags were used, which must have made sense for the evaluator but turned out to be too ambiguous. For example, we found several comments like "translation is a little bit literal," "too literal" or simply saying "good style," or even "great style!", thus lacking enough references to conclude whether the translation needs edition.

From each filtered collected translation, a number of different linguistic measures are extracted and converted into an array of numerical data, paired with the three evaluations it received, which work as the label. The automatic evaluator is trained using these feature arrays and their labels. The model thus created will be used to predict evaluations for new translations.

Peer review through Open EdX has a time constraint. If the window of time given to provide reviews is too small, some participants may be left out. If it is too big, participants submitting on Monday may not be there on Friday to assess the work that one of their peers submitted in the last moment. That is the reason why we have only collected translations with three peer assessments.

Extracted information

The idea behind the evaluation tool is to rely on the abundance of data that can be gathered through an MOOC in order to build a robust prediction model. This model is the building block of the evaluation tool and as such, it should provide evaluations as immediate as possible, so that a student can submit a translation and promptly receive feedback on it. The most accurate evaluators, however, are those provided by features that are very costly to extract from raw text, like language models, parse trees, or language pair information. Extracting these features is costly both in terms of time and in expert knowledge.

The features extracted from the translations have been chosen to minimize computational overhead. That is why, we avoided to use features extracted by other systems, like Moses translation models (Koehn et al. 2007), or language models (Stolcke 2002), or any other resource that would make the evaluation asynchronous, since we aim to provide quick feedback to the user. The features are a combination of some baseline features mentioned by Specia (precisely those that can be straightforwardly extracted), some features taken from the field of forensic linguistics and authorship attribution (García-Barrero 2012), TF-IDF word counts in relation with the whole corpus of collected translations,[5] and basic POS tagging, for a total of 74 features. The target labels were created by simply calculating the arithmetic mean between the three assessments, as shown in Figures 2 and 3 for the two texts.

Modeling methods

As a baseline method, we considered both standard and ridge regression for translation quality assessment having as target the average of the evaluations, as done by (Wisniewski, Singh, and Yvon 2013). The former method, however, could not be run because the number of predictors exceeds the number of observations, and the latter gave very poor results. Given the challenging

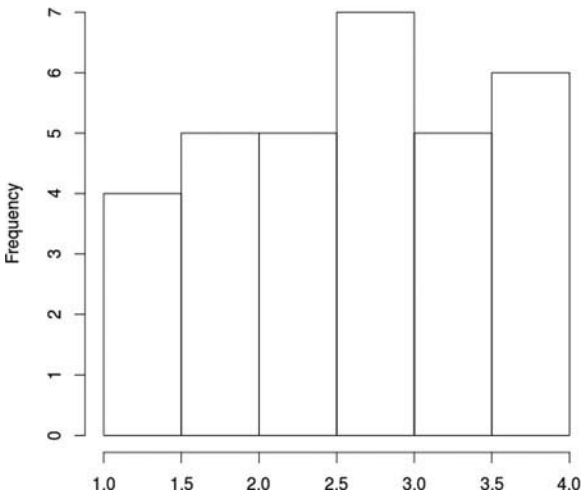

Figure 2. Histogram of average quality assessment (Text1).

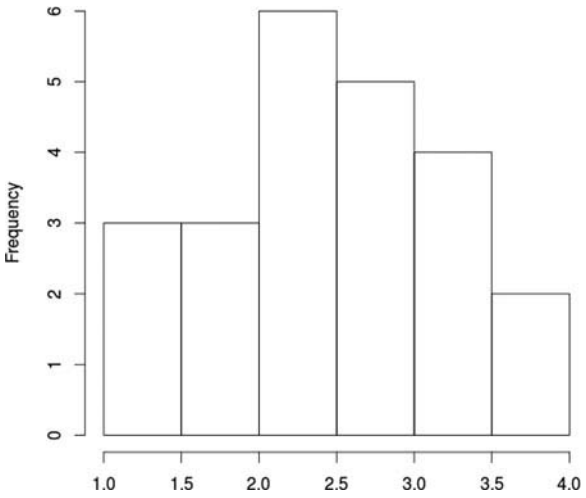

Figure 3. Histogram of average quality assessment (Text2).

difficulty of the problem, we considered two state-of-the-art machine learning methods: a Random Forest (RF) (Breiman 2001) and a Relevance Vector Machine (Tipping 2001), both developed using the R language for statistical computing (R Development Core Team 2016).

The RF is an ensemble approach that consists of a set of randomized decision trees. The main principle behind ensemble methods is that a group of "weak learners" can come together to form a strong learner. In an RF, the weak learner is a decision tree. The parameters are the number of trees and the number of variables tried at each split. The method is very fast to train, and able to deal with unbalanced and missing data. The RF is also able to give a precise estimation of prediction error—called the Out-of-Bag(OOB)

error. The number of variables m tried at each split is set to one third of the total number of variables ($m = 24$, in our case), as is standard practice. We tested a varying number of trees in the sequence 10^1 to 10^3 in steps of $10^{0.1}$, guided by the OOB. For each number of trees, 50 repetitions were run, and their average OOB was recorded.

The Relevance Vector Machine (RVM) is a sparse Bayesian method for training the generalized linear models which has the same functional form as the support vector machine for regression (SVM-R). It is a kernel-based technique that typically leads to sparser models than the SVM-R, and may also perform better in many cases. The RVM introduces a prior over the weights governed by a set of hyperparameters, one associated with each weight, whose most probable values are iteratively estimated from the data. The RVM has a reduced sensitivity to hyperparameter settings than the SVM-R. For the RVM, we choose the RBF kernel and optimize the inverse kernel width parameter γ. Theoretically, the whole training set could be used to fit the RVM—without cross-validation. However, resampling is still needed to choose the best value for γ; therefore, a 10×10 CV procedure (10 times 10-fold cross-validation) is used to evaluate predictive performance using values in the sequence 10^{-3} to $10^{3.5}$ in steps of $10^{0.01}$.

Results and discussion

For Text1, the best result was achieved with 1000 trees (the maximum number tested), with a predictive mean of squared residuals (MSR) equal to 0.503, corresponding to a 33.8% of explained variance or $R2$ coefficient. The best result with the RVM was achieved with $\gamma = 4.9 \times 10^{-4}$, with a predictive MSR equal to 0.414, corresponding to a 45.5% of explained variance. In order to put this result in context, we also report the mean absolute error(MAE), a quantity that is more amenable to interpretation and recommended over the MSR by several authors—see (Willmott and Matsuura 2005). The computed predictive MAE for the RVM model is 0.516 which means that, on an average, the quality is wrongly predicted by a factor of one-half (e.g., quality of 3.5 instead of 3).

For Text2, the best result was again achieved with 1000 trees, with a predictive MSR equal to 0.377, corresponding to a 28.5% of explained variance. The best result with the RVM was achieved with $\gamma = 2.0 \times 10^{-4}$, with a predictive MSR equal to 0.375, corresponding to a 28.8% of explained variance. In this second experiment, although the predictive errors are better, the variance of the quality was much higher—as seen in Figure 3—which leads to lower percentages of explained variance. The predictive MAE for the RVM model is 0.478.

One reason behind the large variability in quality evaluations is that some translations may have had bad evaluations due to translation errors—making them differ very little from good translations, but still in need of edition for

them to be valid—thus getting bad evaluation from some peer reviewers, but not from others.

Human evaluations also differ by their inherent subjective character; consider, for instance, the case of a typical translation, getting evaluations such as {4,3,3} or {1,2,2}, thus avoiding mixing 1 and 4 (a difference of 3 points between marks). Very seldom the difference between marks reached 2 points, and very often there were differences of 1 point in the evaluations. This means that the boundaries between adjacent evaluations are blurry, even for the human assessment.

Concerning the characteristics that are most important for determining the quality of a translation, it is really difficult to think in terms of individual features that could discriminate well in most situations. For instance, the slight difference between the sentences: "*Abdelkrim El Khattabi fought alongside the Spaniards*" vs. "*Abdelkrim El Khattabi fought against the Spaniards*" is a difficult problem to solve by a computer, since meaning is the key. In our study, on analyzing the results delivered by both the RFs and the RVMs, we found that the most discriminant features were those related to *n*-grams; then the part-of-speech tagging (frequency of nouns, verbs, adverbs, adjectives) and then the TF-IDF for a given word in relation to the corpus of translations, i.e. whether a given word of *n* characters (*n* from 3 to 10) present in the translation are also found in many other translations.

Conclusions

Embedding machine assistance in a translation framework targeting trainees opens new possibilities for the improvement of both human and machine translation.

The main idea behind the present work is to test whether a tool that provides automatic evaluation of translation quality could be successfully implemented within an MOOC. In the process of learning a new language, the ability to get immediate feedback on a given translation could be a helpful step, much needed to develop courses on second language acquisition, or courses about translation like the one set up for this work. In particular, the use of MOOC-based peer-review tools as developed for this work could allow university departments or official language schools to easily build all kinds of language corpora, given that they have a steady base of students to participate, and their official backup would attract more participants. With appropriate training of evaluation rubrics, participants would provide valuable annotations for those corpora, which in turn could help gain valuable insight on language acquisition or translation.

An MOOC using an automatic tool for translation quality evaluation could gather huge amounts of significant data in the field of translation studies. This means that a feedback loop could be easily established—where

new translations help refine the tool—opening paths to explore translation in unprecedented ways. Parallel corpora could be then gathered, with a focus on the subjects chosen by the course creator. Corpora of translations created by both native and non-native speakers could be also gathered, as well as corpora of translations created by learners in different stages of learning. All this information could help to identify common mistakes and create new language learning content, usable both in online and face-to-face environments.

Acknowledgments

This work is supported by the 7th Framework Program of the European Commission through the International Outgoing Fellowship Marie Curie Action (IMTraP-2011-29951) and also by the Spanish Ministerio de Economía y Competitividad and European Regional Development Fund, through the postdoctoral senior grant Ramón y Cajal and the contract TEC2015-69266-P (MINECO/FEDER, UE).

Notes

1. https://tradares.wordpress.com
2. The course is already closed, but there is a backup that would allow to put it back online anytime on request.
3. http://ucatx.cat/
4. https://www.researchgate.net/publication/283053235_CorpusTRADARES-2015
5. In such a way that, if a given word in a new translation is also found in the rest of translations, it will count toward a positive evaluation.

References

Balfour, S. P. 2013. Assessing writing in MOOCs: Automated essay scoring and calibrated peer review. *Research and Practice in Assessment* 8 (1):40–48.

Breiman, L. 2001. Random forests. *Machine Learning* 45 (1):5–32. doi:10.1023/A:1010933404324.

Christensen, G., A. Steinmetz, B. Alcorn, A. Bennett, D. Woods, and E. J. Emanuel. 2013. The MOOC phenomenon: Who takes massive open online courses and why? SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2350964.

García-Barrero, D. 2012. La atribución forense de autora de textos en árabe estándar moderno. Estudio preliminar sobre el potencial discriminatorio de palabras y elementos funcionales. Ph.D. thesis, University of Granada.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, and C. Dyer. 2007. Moses: Open source toolkit for statistical machine translation. Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions, Prague, 177–80.

Pappano, L. 2012. *The year of the MOOC*. The New York Times. The New York Times Company. Date: 2012/11/04.

R Development Core Team. 2016. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Shah, K., T. Cohn, and L. Specia. 2015. A bayesian non-linear method for feature selection in machine translation quality estimation. *Machine Translation* 29 (2):101–25. doi:10.1007/s10590-014-9164-x.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. Proceedings of the Association for Machine Translation in the Americas, Cambridge, Massachusetts, 223–31.

Specia, L., M. Turchi, N. Cancedda, M. Dymetman, and N. Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. Proceedings of the 13th Conference of the European Association for Machine Translation, Barcelona, 28–37.

Stolcke, A. 2002. Srilm-an extensible language modeling toolkit. Proceedings of the International Conference on Spoken Language Processing, Denver, 257–86.

Tipping, M. E. 2001. Sparse bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research* 1:211–44.

Willmott, C. J., and K. Matsuura. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* 30 (1):79. doi:10.3354/cr030079.

Wisniewski, G., A. K. Singh, and F. Yvon. 2013. Quality estimation for machine translation: Some lessons learned. *Machine Translation* 27 (3–4):213–38. doi:10.1007/s10590-013-9141-9.