Parallel Construction: A Parallel Corpus Approach for Automatic Question Generation in Non-English Languages

Benny G. Johnson[0000-0003-4267-9608], Jeffrey S. Dittel, Rachel Van Campenhout[0000-0001-8404-6513], Rodrigo Bistolfi, Aida Maeda, and Bill Jerome

VitalSource Technologies, Pittsburgh PA 15218, USA
benny.johnson@vitalsource.com

Abstract. Automatic question generation (AQG) has many diverse applications in educational contexts. To bring these benefits to as many students as possible, it is prudent to expand AQG capabilities in as many languages as possible. However, English remains the dominant language in AQG research, and the required natural language processing tools for other languages are often under-resourced relative to English, which can make developing AQG pipelines difficult or impractical altogether. An approach called *parallel construction* has been developed to leverage existing English AQG systems for AQG in other languages. The benefits of this parallel construction approach are described, and examples of questions generated from Spanish and Brazilian Portuguese textbooks using the parallel construction method are presented and discussed.

Keywords: Textbooks, Learn by doing, Automatic question generation, Machine translation, Parallel corpus methods, Parallel construction.

1 Introduction

Formative practice is an established learning technique used in many educational contexts and known to benefit all students, but is especially useful for struggling students [2]. Formative practice acts as no-stakes practice testing; students answer questions meant to foster learning and prepare them for high-stakes assessments without the worry of being graded for their responses. This learn by doing method of studying can be causal to learning [11] and is especially helpful in digital learning environments that offer immediate feedback to students [6].

Automatic question generation (AQG) using natural language processing (NLP) can scale creation of questions from textbook content in a way that is unattainable through human effort. AQG is a popular area of research in education given the multitude of application possibilities across subjects, ages, and learning and assessment approaches [12]. Recent research on AQG applied as formative practice to natural learning contexts in English has shown that automatically generated (AG) questions can achieve performance equivalent to human-authored questions on metrics of engagement, difficulty, persistence [17, 18] and discrimination [10]. The established performance and future potential of AG questions confirm they should be made available for as many students as possible, including learners in languages other than English. Of the AQG systems included in a recent systematic review [12, Table 14], only 12 of 72 were for non-English languages (Chinese, Japanese, Indonesian, Thai, and Punjabi). The authors noted an increase in publications on AQG in non-English languages relative to a previous review, speculating that interest in generating questions in other languages could be due to increased interest in NLP research in those languages. For the languages in the current work, Spanish and Brazilian Portuguese, no AQG systems are reported in [12], while another recent review [4] includes one system for Portuguese. More recent research in these languages reports translation of an adapted SQuAD data set [5] to Spanish for use in AQG [15], and systems for factual question generation in Portuguese [8, 14]. Still, English remains by far the dominant focus of AQG research.

It would be highly desirable to have a way to leverage the benefits of English AQG research and the substantial development effort that has gone into English AQG systems when working with content in other languages. AQG systems are typically quite complex, involving a variety of NLP methods and tools, such as part-of-speech tagging, dependency parsing, and vector space embedding. For AQG in English, sufficiently robust and accurate NLP tools are readily available. However, even though several AQG systems in non-English languages have been reported [12], implementing similarly robust AQG pipelines in other languages can be problematic because the NLP tools are often under-resourced relative to English, sometimes significantly so. Consequently, several capabilities needed in AQG are often not as performant for lower-resource languages, meaning that achieving sufficiently high reliability may not always be possible. When this is the case, it is not practical to build an AQG pipeline directly in the source language. Furthermore, even for a language like Spanish (the fourth most-spoken language in the world) where NLP technology is not under-resourced to the degree that many other languages are, there are still far more AQG systems for English. The ability to reuse these existing systems could save a considerable amount of development and empirical validation work, helping to expand AQG in other languages.

This paper presents a method called *parallel construction*, intended as a complementary approach to implementing AQG directly in the source language. Parallel construction uses machine translation (MT) and a parallel corpus approach to enable an English language processing pipeline to be used for AQG in other languages. Rather than simply back-translating generated English questions, which would be inadequate due to the large gap in quality that still exists between human translations and MT [9], MT is instead used to create a parallel corpus from the original text, and then parallel corpus techniques are used to construct the source language questions directly with the requisite fidelity. Parallel corpus methods [13, 19] enable knowledge about text in one language to be leveraged for tasks in another language by making use of alignment information for documents, sentences, and words. For AQG, this is realized in two important and complementary ways. First, the results of the NLP analysis of the English text can be applied to AQG in the source language as well through the alignment, even when sufficiently accurate NLP tools are not available for the source language. Second, the alignment information enables a source language version of the English questions to be constructed directly from the original source language text, which was authored by a human subject matter expert and has much higher linguistic quality. This sidesteps the quality issues for MT-generated text that make translation-of-a-translation approaches unacceptable. Therefore, a parallel corpus formulation of AQG enables a way to have the best of both worlds by exploiting the relative advantages of each version of the text.

Furthermore, no AQG decisions (content, pedagogical, or otherwise) need to be implemented using NLP on the source language, which again, is not always practical. Instead, the decisions made in English can be used to drive question construction in both languages. How and why these decisions are made are immaterial to parallel construction, thereby enabling broad application of the method. The parallel construction process mirrors the textual manipulations made in English, by applying their appropriately localized equivalents directly to the aligned source language content, thereby reusing the knowledge base built into the English AQG system.

In the remainder of the paper, the parallel construction method is described in detail and applications to question generation for Spanish and Brazilian Portuguese textbooks are presented and discussed.

2 Methods

2.1 Parallel Corpus Approach

Parallel corpus-based approaches [13, 19] can be used to address a diverse array of NLP problems such as construction of bilingual dictionaries, cross-language information retrieval, and MT itself. A centrally important concept in parallel corpus methods is alignment. This means finding the sentences that correspond to each other in the original text and the translated version, and then identifying the corresponding words within those aligned sentences. This is nontrivial, compounded by the fact that sentence and word correspondences are not always one-to-one. In addition, differences in word ordering within corresponding sentences and any MT errors add complexity.

Google Translate was used to create an English version of the source language corpus, as it is a readily available state-of-the-art MT system. For convenience, sentence alignment was achieved by tokenizing the source language corpus into sentences and sending them to the translation service one at a time. For textbook content, the focus of this work, this can be performed more reliably than with arbitrary text. For word alignment, the best methods are in general statistically based [16]. Here, the `fast_align` method [7], an efficient reparameterization of IBM Model 2 for statistical machine translation, was used.

2.2 Parallel Construction for AQG

The parallel construction method is applicable to template-based, rule-based, and some statistical approaches to AQG, which are the most common procedures of transformation [12]. For illustration, AQG will be described in terms of a rule-based expert system, which is the type of AQG system used in the present work. The system's production rules (rules of the form condition ⇒ action) make the decisions of the AQG strategies and carry out the individual steps of question construction. Notably, it is typically the rules' applicability conditions, not their actions, that require sophisticated NLP analyses, and the productions themselves (e.g., the content transformations) are much more straightforward. The parallel construction method is aware of all possible productions the rules can make so that they can be implemented equivalently (localized) for the source language. However, parallel construction does not require localization of the production rules themselves; there are no analogous rules for the source language involved, and thus analogous NLP capabilities for the source language are not needed.

The first step in AQG is selection of content knowledge from which a question will be made. A common example is a single sentence from a textbook (which will be used for ease of illustration in the examples to follow), but could also be several sentences (such as a paragraph), or another type of content altogether, like a glossary entry. Suppose the English AQG system decides to select the sentence "This is a good sentence for creating a question," for transformation into a question. The parallel construction process then finds the corresponding Spanish sentence "Esta es una buena oración para crear una pregunta," in the original text using the sentence alignment information. We thus see how content knowledge selection can be mirrored in Spanish without attempting to replicate the corresponding decision-making logic using source language NLP, which might not be feasible. Instead, knowledge obtained in one language is applied to facilitate a task in another language, which is the essence of a parallel corpus approach.

The overarching strategy of parallel construction is as follows. The English AQG system operates on the translated text exactly as usual. The system makes step-by-step decisions according to the details of its AQG algorithms. Each decision can cause one or more manipulations to be made to a sample of text, which can be a subset of the English corpus or the output of previous manipulation step(s). The entire sequence of decisions and associated manipulations leads from the input English text corpus to the output English questions. In parallel construction, a process is run side-by-side with the English AQG. Every time a manipulation is applied to English text, the equivalent manipulation is carried out on the corresponding source language text using the sentence and word alignments. In this way, by the time the English questions are fully developed, they are also fully developed in the source language because they are always kept up to date in parallel. Notably, knowledge of the AQG decisions is not needed by the parallel construction process, only the manipulations that need to be made as a result.

Given an English AQG system, parallel construction requires much less development effort compared to direct construction in the source language, as it makes most of the details of the English AQG implementation irrelevant.

3 Application

3.1 AQG in English

This parallel corpus approach was applied to an existing AQG system [18] that was originally built for generating questions from English language textbooks. The two types of questions included in the examples below are matching and fill-in-the-blank (FITB) cloze questions. It is important to note that the AQG system used in this work has been well-studied, with its performance on several key metrics characterized [10, 17, 18]. When this is the case, it also provides relevant information about the questions that will be produced by parallel construction, and as such is another important dimension of reuse. The only additional potential source of error is from the parallel construction process itself, which as will be seen is very low.

3.2 Example 1: AQG in Spanish

Parallel construction AQG was run on a Spanish-language macroeconomics textbook [1]. There were 684 questions generated from the English translation of the textbook, of which 632 (92.4%) were able to be created in Spanish through parallel construction. Cases in which parallel construction cannot be carried out, which account for the construction rate of less than 100%, are discussed below. Step-by-step generation of a matching question from a sentence on page 190 of the textbook is shown in Table 1. Steps in English are denoted by 1(eng), 2(eng), …, and parallel steps in Spanish by 1(spa), 2(spa), …, etc. All English steps are created by the AQG system's production rules; all Spanish steps are created from the English steps using parallel construction.

As in every case, all AQG decisions are made by the English-language system; no decisions involve NLP in the source language. For example, the original Spanish version of the sentence was located using sentence alignment after it was selected in English, not based on direct analysis of its suitability. Parallel construction needs no knowledge of the AQG decisions, only the actions that result. To underscore this property, the decision-making logic of the English production rules is deliberately omitted.

Table 1. Parallel construction steps for a matching question in Spanish.

Step	Description	Output
1(eng)	A production rule in the English AQG system selects a sentence for question generation:	However, during the 1980s many borrowing LDCs were unable to cope with the burden of their foreign debt - a situation known as the LDC debt crisis - and, perhaps as a consequence, their economic growth. countries experienced a serious decline.
1(spa)	The corresponding Spanish sentence is retrieved using the sentence alignment:	Sin embargo, durante la década de 1980 muchos PMD prestatarios no pudieron hacer frente a la carga de su deuda exterior –situación que se conoce con el nombre de crisis de la deuda de los PMD– y, quizá como consecuencia, el crecimiento económico de estos países experimentó una grave disminución.
2(eng)	Additional production rules in the English system select the answer words:	borrowing, crisis, decline

2(spa)	The corresponding Spanish words are retrieved using the word alignment:	prestatarios, crisis, disminución
3(eng)	The final English question is constructed as follows (alphabetizing choices):	However, during the 1980s many ______ LDCs were unable to cope with the burden of their foreign debt - a situation known as the LDC debt ______ - and, perhaps as a consequence, their economic growth. countries experienced a serious ______. Choices: borrowing, crisis, decline
3(spa)	The final parallel question in Spanish is:	Sin embargo, durante la década de 1980 muchos PMD ______ no pudieron hacer frente a la carga de su deuda exterior –situación que se conoce con el nombre de ______ de la deuda de los PMD– y, quizá como consecuencia, el crecimiento económico de estos países experimentó una grave ______. Opciones: crisis, disminución, prestatarios

The English sentence illustrates the noise that can happen with MT. Near the end there is a syntax error "...growth. countries..." Not only is the meaning difficult to discern here, it is not entirely faithful to the Spanish source text, which says the economic growth of the countries experienced a decline, not the countries themselves, as would be one possible reading of the English text. The poor linguistic quality of the English text did not prevent AQG from succeeding. However, the translated sentence would never be included in an English textbook as is, nor is the resulting English question acceptable for students. Despite this, the Spanish question produced by parallel construction is entirely acceptable, having the same linguistic quality as the original source language text. This is due to parallel construction operating on the original text directly.

By contrast, compare the final question in Table 1 to the result of merely back-translating the English question to Spanish:

> Sin embargo, durante la década de 1980, muchos PMA ______ no pudieron hacer frente a la carga de su deuda externa, una situación conocida como la ______ de la deuda de los PMA, y, tal vez, como consecuencia, su crecimiento económico. Los países experimentaron un grave ______.
> Opciones: crisis, declive, prestatarios

The difference is stark. This question is of much lower linguistic quality than the one obtained by parallel construction. It retains the original MT error, thereby making the Spanish version unacceptable as well. Also note that the acronym "PMD" in the original Spanish content, which stands for "países menos desarrollados" (translated to English as "LDC" = "less developed countries"), becomes "PMA" upon back-translation, which is "paises menos avanzados." This translation is actually a correct one, but the question would be problematic for students because it introduces a departure from the textbook's notation without explanation. Therefore, while that translation would likely be acceptable in many circumstances, for educational applications it is not. The parallel construction method is not susceptible to this problem.

It is important to note that the word alignment for this sentence was not perfect; not all English words were able to be mapped, caused at least in part by the MT noise present. However, in this case the imperfect alignment does not compromise parallel construction since the subset of words that are relevant was mapped correctly. Although incomplete or incorrect alignment of the answer words themselves would have been problematic, this example shows that the method is able in some cases to be robust against alignment errors and still succeed despite them.

What if alignment had in fact failed on words that were required by parallel construction? This could happen in at least two ways. First, if the required words were unable to be aligned it is not possible to carry out the parallel step. When this happens, or if for any reason the step cannot be performed, the question can simply be discarded. This typically has resulted in less than 10% of questions generated in English being discarded. Second, if the word alignment is incorrect, the system still has the potential to produce a valid question, but one that is not identical to its English counterpart. While this is not ideal, it nonetheless mitigates the risk of errors in meaning or dysfluency that can happen with back-translation, since the source language question will still be accurate and the linguistic quality of the source text will be preserved.

3.3 Example 2: AQG in Brazilian Portuguese

Here, parallel construction AQG was run on a Brazilian Portuguese-language psychopathology textbook [3]. There were 969 questions generated in English, with 942 (97.2%) questions in Portuguese created. A representative FITB question, from a sentence on page 436 of the textbook, is shown in Table 2.

Table 2. Parallel construction steps for a FITB question in Portuguese.

Step	Description	Output
1(eng)	A production rule selects an English sentence for question generation:	Phenomena of the autonomic nervous system (sympathetic and parasympathetic) can occur, such as sweating profusely, presenting fever, tachycardia and tremors, sometimes gross (including flapping, or asterisks).
1(por)	The corresponding Portuguese sentence is retrieved using the sentence alignment:	Podem ocorrer fenômenos do sistema nervoso autonômico (simpático e parassimpático), como suar profusamente, apresentar febre, taquicardia e tremores, às vezes grosseiros (inclusive flapping, ou asteríxis).
2(eng)	A production rule selects the English answer word:	autonomic
2(por)	The corresponding Portuguese word is retrieved using the word alignment:	autonômico

3(eng)	The final English question is constructed as follows:	Phenomena of the ______ nervous system (sympathetic and parasympathetic) can occur, such as sweating profusely, presenting fever, tachycardia and tremors, sometimes gross (including flapping, or asterisks).
3(por)	The final parallel question in Portuguese is:	Podem ocorrer fenômenos do sistema nervoso ______ (simpático e parassimpático), como suar profusamente, apresentar febre, taquicardia e tremores, às vezes grosseiros (inclusive flapping, ou asteríxis).

Note that the selected English sentence contains a translation error: the medical term “asteríxis” is mistranslated as “asterisks.” While this results in a corrupted English question being generated, the Portuguese question is still correct despite this significant error since parallel construction works directly on the original Portuguese text.

Suppose the mistranslated word “asterisks” had been selected as the answer for the English question, instead of “autonomic.” In this case, it turns out that “asterisks” was not able to be aligned to a source Portuguese word; this was likely a consequence of the translation error. This would make Step 2(por) unable to be performed and result in the question being discarded. Therefore, the MT error would still not lead to an erroneous question in Portuguese, although back-translation would.

4 Conclusion

We have presented parallel construction as an approach to AQG in non-English languages, specifically as an alternative to AQG directly in the source language. The parallel construction method involves creating a source language-English parallel corpus using MT, aligning that corpus, generating questions using an English AQG system, and applying the results of the English AQG process in parallel to construct the corresponding questions from the source text. In this way, the knowledge base and development effort that went into the English AQG system are reused, while the questions produced have the linguistic quality of the source text.

A major advantage of parallel construction is it is largely independent of the implementation details of the English AQG system, making it broadly applicable. As seen in the examples provided, it is also robust to errors and noise that can occur during MT. Parallel construction is also applicable to many other question types than those presented. We are currently extending the implementation to question types such as multiple choice and wh-questions that are already generated by our English AQG system.

The next major step is empirical evaluation of the questions generated through parallel construction. An evaluation by subject matter experts teaching in Spanish has been conducted in several subject domains and results will be reported in future work.

References

1. Abel, A. B., & Bernanke, B. S. (2004). *Macroeconomía* (4th ed.). Madrid: Pearson Educación.
2. Black, P., & William, D. (2010). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan, 92*(1), 81-90. https://doi.org/10.1177/003172171009200119
3. Dalgalarrondo, P. (2019). *Psicopatologia e semiologia dos transtornos mentais* (3rd ed.). Porto Alegre: Artmed.
4. Das, B., Majumder, M., Phadikar, S., & Sekh, A. A. (2021). Automatic question generation and answer assessment: a survey. *Research and Practice in Technology Enhanced Learning*, *16*(1), 1-15.
5. Du, X., Shao, J., & Cardie, C. (2017). Learning to ask: neural question generation for reading comprehension. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1342-1352). https://doi.org/10.48550/arXiv.1705.00106
6. Dunlosky, J., Rawson, K., Marsh, E., Nathan, M., & Willingham, D. (2013). Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4-58. https://doi.org/10.1177/1529100612453266
7. Dyer, C., Chahuneau, V., & Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. *NAACL HLT 2013 - 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Main Conference*, *June*, 644–648.
8. Ferreira, J., Rodrigues, R., & Gonçalo Oliveira, H. (2020). Assessing factoid question-answer generation for Portuguese. *Proceedings of the 9th Symposium on Languages, Applications and Technologies (SLATE 2020).* Schloss Dagstuhl-Leibniz-Zentrum für Informatik. https://doi.org/10.4230/OASIcs.SLATE.2020.16
9. Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, *9*, 1460-1474. https://doi.org/10.1162/tacl_a_00437
10. Johnson, B. G., Dittel, J. S., Van Campenhout, R., & Jerome, B. (2022). Discrimination of automatically generated questions used as formative practice. *Proceedings of the Ninth ACM Conference on Learning@Scale*. https://doi.org/10.1145/3491140.3528323
11. Koedinger, K. R., McLaughlin, E. A., Jia, J. Z., & Bier, N. L. (2016, April). Is the doer effect a causal relationship? How can we tell and why it's important. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 388-397). Edinburgh, United Kingdom. http://dx.doi.org/10.1145/2883851.2883957
12. Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, *30*(1), 121-204. https://doi.org/10.1007/s40593-019-00186-y
13. Lefer, M.-A. (2020) Parallel corpora. In M. Paquot & S. T. Gries (Eds.), *A practical handbook of corpus linguistics*. Springer, Cham. https://doi.org/10.1007/978-3-030-46216-1_12
14. Leite, B., Cardoso, H. L., Reis, L. P., & Soares, C. (2020). Factual question generation for the Portuguese language. *Proceedings of the 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)* (pp. 1-7). http://dx.doi.org/10.1109/inista49547.2020.9194631
15. Mamani Maquera, F., Paz Valderrama, A., & Castro Gutierrez, E. (2019). Performance evaluation of recurrent neural network on large-scale translated dataset for question generation

in NLP for educational purposes. *Proceedings of the 17th LACCEI International Multi-Conference for Engineering, Education, and Technology*. http://dx.doi.org/10.18687/LACCEI2019.1.1.178

16. Santos, A. (2011). A survey on parallel corpora alignment. *Proceedings of MI-Star 2011*, 117-128.
17. Van Campenhout, R., Brown, N., Jerome, B., Dittel, J. S., & Johnson, B. G. (2021). Toward effective courseware at scale: investigating automatically generated questions as formative practice. *Proceedings of the Eighth ACM Conference on Learning@Scale* (pp. 295-298). https://doi.org/10.1145/3430895.3460162
18. Van Campenhout, R., Dittel, J. S., Jerome, B., & Johnson, B. G. (2021). Transforming textbooks into learning by doing environments: an evaluation of textbook-based automatic question generation. *Third Workshop on Intelligent Textbooks at the 22nd International Conference on Artificial Intelligence in Education. CEUR Workshop Proceedings, ISSN 1613-0073* (pp. 60–73). http://ceur-ws.org/Vol-2895/paper06.pdf
19. Véronis, J. (2000) From the Rosetta stone to the information society. In J. Véronis (Ed.), *Parallel text processing* (pp. 1-24). Springer, Dordrecht. https://doi.org/10.1007/978-94-017-2535-4_1