Automated question answering via natural language sentence similarity: Achievements for Brazilian e-commerce platforms

Víctor Jesús Sotelo Chico[1], Luiz Zucchi[1,2], Daniel Ferragut[1,2], Rodrigo Caus[2], Victor Hochgreb de Freitas[2], Julio Cesar dos Reis[1]

[1]Institute of Computing – University of Campinas, Campinas, Brazil

[2]GoBots Company, Brazil

jreis@ic.unicamp.br,{rodrigo.caus,victor}@gobots.com.br

{victor.sot.c,luizeduardoaraujozucchi,danielpferragut}@gmail.com

Abstract. *Chatbots have become indispensable for answering e-commerce customer queries, which is crucial for selling products online. However, in Brazilian e-commerce, finding scalable solutions can be challenging. This article proposes an automatic question-answering system by replying to incoming questions with Frequently Asked Questions from stores. Our solution builds a store-specific database populated with question-answer pairs by generating the embedding of questions. We rank candidate questions using a neural network to retrieve known answers. Our solution was tested on data from South American e-commerce platforms in Portuguese and Spanish. The development approach achieved* 97.75% *of satisfaction with the given answers.*

1. Introduction

Customer service has become an essential part of e-commerce. Offering an excellent experience to clients through a virtual platform is the key to establishing trust. Customers spend hours researching products with thousands of choices available, and supporting eventual questions in real time about the product is vital to guarantee a sale.

E-commerce companies focus on automatizing this process with artificial intelligence solutions; some solutions involve finding the question's *intent* (the user's purpose) and its *entities* (relevant terms and objects in the query). This approach is hard to scale because answering a question requires knowing all possible intents and entities for that type of product. Furthermore, while large language models can provide answers for e-commerce stores, they are often private and too expensive for small businesses.

This article offers a solution for customer questions by comparing them to previously answered ones. If a past customer has asked a question about a product before, our solution uses the answer provided by customer service. The defined approach does not rely on detecting any *intent* or *entities* for a product or an incoming question in free text.

We face a challenging task in sentence similarity detection. First, it requires a coherent way to represent sentences so that questions presenting similar meanings also have similar numeric values. In this context, the main issue is to obtain models for our specific goal, which must address the informal Portuguese language spoken on e-commerce platforms. Although we may have accurate sentence representations, finding similar questions takes time. Our problem requires handling a high precision rate, as a misleading answer can be an inadequate experience for the customer.

A semantic search retrieval process finds questions already answered that have the same meaning as the new incoming question. This first stage acts as a filter to get possible similar candidates. Our solution converts sentences to vectors to enable a similarity computation and verifies that they are semantically identical; we perform this conversion with the Universal Sentence Encoder model large multilingual [Cer et al. 2018]. Then a semantic search occurs in a database using Elasticsearch [Gormley and Tong 2015]. This reduces the vector space. Finally, our trained neural network performs a classification task to evaluate pairs of questions' similarity in these reduced spaces to obtain the sentence with the highest score (similar).

We deployed the system to online stores in the massive marketplace in Brazil and monitored the system for one month. We report results from May 14, 2023, to June 14, 2023. Our system helped to answer 13,991 input questions from real-world customers.

The evaluation results show that our system can answer questions that current chatbots have difficulties, increasing the total number of questions answered correctly, and it is suited to improving customers' experience without needing manual labor.

The remainder of this article is organized as follows: Section 2 presents related studies. Section 3 reports on the full description of the developed system. Section 4 describes the experimental results in deploying and assessing the solution in the *GoBots* company environment, along with an automatic evaluation with annotated data. Section 5 discusses the obtained findings; Section 6 presents the conclusion and future work.

2. Related Work

Question-Answering systems are well-known Natural Language Processing applications; in literature [Kulkarni et al. 2019] propose a solution that analyses the product's description and user's review content by exploring semantic annotation based on ontologies, an intent classifier, and an answer ranking component.

[Chen et al. 2019] use sentence embedding with two multi-layer convolution networks: one to find relevant user review snippets and another to obtain the answer for the desired question from these snippets. [Gupta et al. 2019] that follows a similar with a review-based QA to synthesize the review and answer the customer questions. The solutions provided rely on user reviews and community answers, which may not always be entirely reliable due to their bias towards users' opinions. The closest study to our proposal is the one built by [Sakata et al. 2019]. In their work, a suitable response is chosen based on the query-question similarity through a query and frequently asked question chart and their answers. Furthermore, they consider the query-answer relevance with a BERT [Devlin et al. 2019] based component. This solution produced relevant results.

However, the research focuses on a formal domain in contrast e-commerce context in marketplace platforms that handle several online stores. Besides, for the query-question component, a system focused on Japanese queries (*TSUBAKI* [Shinzato et al. 2008]) instead of universal ones was explored. A problem with this solution is that it requires an expensive GPU server to train and run, which can be a problem for smaller companies.

Moreover, [Mass et al. 2020] study FAQ retrieval using BERT to train two models to match *questions* and *answers* relevance from a given query. Furthermore, the authors proposed using question paraphrasing to overcome the size limitation of datasets.

[Gupta and Carvalho 2019] propose a multiple deep learning architecture based on attention [Vaswani et al. 2017] mechanism to compute both query-question and query-answer similarities to retrieval a response.

Alternatively, [Finardi et al. 2021] explore FAQ retrieval oriented to a specific domain constructing BERTaú, a Portuguese financial model, which learns a specific context representation. This model uses the data from the Itau Bank chatbots to train a neural network. The research applied the models for FAQ retrieval as a robust test over the objective domain, improving over other BERT models.

Although the existing approaches match question-answer relevance, e-commerce solutions might not benefit because historical stores' answers can be simple short sentences. For example, a simple answer might be *"yes it works"* and customized responses to specific questions such as *"Hi, thanks for buying with us. You can change the product if they have a problem"*. For actual e-commerce applications, this can derive from deploying specific solutions for each store, which is not scalable for big businesses. Thus, e-commerce FAQ retrieval likely finds more benefits from exploring only similarities across questions and criteria for deciding whether to answer.

3. *Deep FAQ*: A question-answering system based on sentence similarity

Our proposed solution called *DeepFAQ* is organized into two parts: *Buildtime* (cf. Subsection 3.1) and *Runtime* (cf. Subsection 3.2). The former consists of operations performed before the system answers a question; the latter consists of operations performed during the execution of the system to find an answer to a new incoming question. The proposed system works alongside another chatbot developed by *GoBots company*. From now on, we refer to this *company* system as the *Base System*. The *Base System* works by extracting intention and keyword entities from a given question and matching patterns between them with pre-made answers. The *DeepFAQ* system works on top of this solution, answering questions that the *Base System* cannot handle. Therefore, the *GoBots* team who manages the bots can control the question that goes to it or any other system parameter.

3.1. Buildtime

Before the system can reply to an incoming question for a given store in an e-commerce platform, it is necessary to generate a vector space from which it is possible to retrieve potential candidate questions. A neural network called *DeepSim*, which ranks questions on their similarity, is trained so that it is possible to use it on the *Runtime* (cf. Subsection 3.2). Figure 1 presents the components and procedures. The upcoming subsections explain each component in detail.

Question Answer retrieval and pre-processing: The first step involves building a vector space with potential candidate questions to retrieve the question-answer pairs from a store. To achieve this, we have created a script that accesses the *Mercado Livre API* and retrieves the questions from the store, obtaining all QA pairs in the store's history since the store joined the *Mercado Livre* e-commerce platform. Once we have the QA pairs from the store, we perform some basic pre-processing: removing stop words, converting the characters to lowercase, and removing accent marks. The process of removing stop words from answers is essential because customer service answers usually contain greetings such as daytime (*e.g.*, good morning, good night, *etc.*). Additionally, there exists a signature from the store's employee.

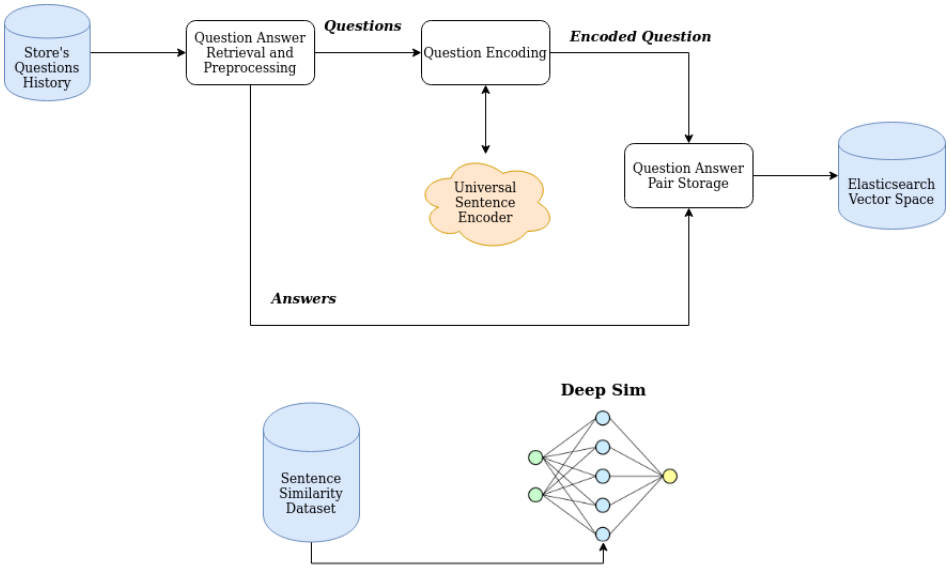

Figure 1. DeepFAQ's Buildtime procedure

Question Encoding: After cleaning QA, we encode these questions using the *Universal Sentence Encoder (USE)* proposed by [Cer et al. 2018]. Encoding involves sending questions to a remote service running to obtain their embedding and associate response. Our solution is suited to replace USE for any other encoders to be updated to improvements over the field of semantic representation.

Question-Answer Pair Storage: After obtaining the sentence embeddings, each QA pair and its question embeddings are sent to our vector space database with *Elasticsearch*. This avoids storing repeated questions so that the candidate retrieval step in *Runtime* is not affected (cf. Subsection 3.2). Finally, we store the QA pair and the question embedding, and the *product identifier* for which the asked question.

DeepSim Training: We trained a neural network called DeepSim alongside the QA pair database population process. We found the optimal architecture for a Siamese network with one layer, 512 neurons, relu activation, and 0.001 learning rate through random search, over this setup we trained using early stopping criteria of three over its loss function. This network compares two encoded questions and outputs the probability of them being identical, improving the system's accuracy for finding similar questions.

Finding Portuguese datasets for sentence similarity for e-commerce can be challenging; the few available, like ASSIN [Fonseca et al. 2016], are small and oriented for formal language, which is unsuitable for e-commerce queries. We decided to combine different kinds of datasets in this scenario. The first kind of data was the *Quora Question Pair* dataset [1], we translated from English to Portuguese using the *AWS Translator* [2].

The second kind of data was auto-generated pairs of sentences based on a pair of sentences template; those templates are common questions. We use a simple script to change selected words from the query to achieve similar or different questions. We ob-

[1] `https://www.kaggle.com/c/quora-question-pairs`

[2] `https://docs.aws.amazon.com/translate/latest/dg/API_TranslateText.html`

tained a dataset from the DeepFAQ system by deploying a beta version to some stores and logging similar questions. We ended up with about 6000 pairs of questions annotated by a freelancer and verified by two others for reliability. All datasets contain two annotated sentences, whether they are similar or not.

3.2. Runtime

Figure 2 presents the system with the active components at the runtime of the solution. The process of answering a new incoming question explores the constructed vector space database (in build-time – cf. subsection 3.1) with the QAs history for a particular online store in the e-commerce platform. Our solution determines if two questions are similar or not. In the following sections, we explain how the system works in more detail.

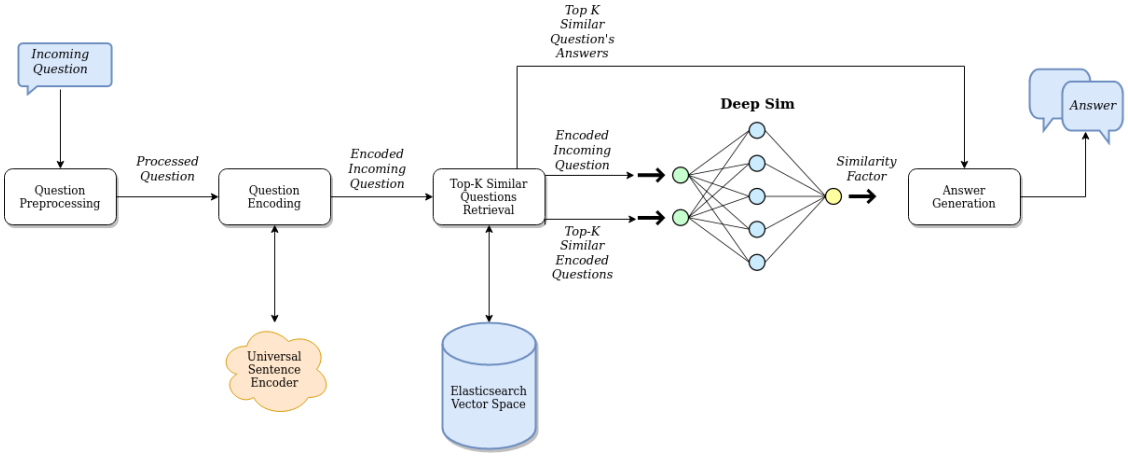

Figure 2. DeepFAQ's Runtime – deployed system ready to answer clients' new incoming NL questions in an online e-commerce store

Question pre-processing and encoding: The basic pre-processing performed on the incoming question is similar to the one made in the *Buildtime*. This processed question is then encoded with the *USE* model, obtaining the sentence embedding.

Top-K similar question retrieval: With the encoded question available, a query is performed in the database holding the vector space of previous questions; this procedure retrieves the top-k most similar questions using cosine similarity, and k is a parameter in the system. The set with the potential candidates holds their encoded questions and their QA pair in natural language. *Elasticsearch* does not implement natively an excellent way to search the top k questions, so we used the *AWS Elasticsearch Service* that is based on *Elasticsearch Open Distro* that comes with a KNN algorithm ready to use.

Ranking candidate questions via the *DeepSim* network: In some tasks, this first retrieval is enough to provide good results to the final user. However, in our case, we still need one more step further to answer questions correctly. The reason for this is that the vectors provided by [Yang et al. 2020], which is a multilingual version of USE encapsulates the intention of a question, but fail to notice that the object (ex. Product).

In our context, this is crucial because a particular question about a product A specification may have a completely different answer than a product B. Therefore, to solve this, each question from the set of top-k candidates is then used as input for the

DeepSim network in conjunction with the incoming encoded questions. This process enables the Deep-sim to rank each element in the top-k candidates by associating them with confidence, reflecting a possibility of being similar to the original incoming question.

Answer generation: Once the candidates are ranked, the solution selects the question candidate in the set with the *highest confidence to being similar*, given by *DeepSim*. The candidate's question must satisfy a probability threshold set by the online store on the e-commerce platform to respond to the client's question, which can change dynamically. Such question has an associate answer pair from the *Store's Questions History* (Figure 1), we use this historical information to provide answers to our clients' questions.

After generating an answer, the system performs post-processing by adding greetings and signatures from the store. Occasionally, responses may contain URLs, and the system checks if they are still available before providing the answer. Finally, if it has a problem, it chooses not to answer. After the system answers, a procedure called answer review is performed by human attendants from the stores to evaluate the answer and ask for its correction.

Continuous populating: When the system fails to answer a question, a human attendant answers the question on the platform as (s)he would do it if there were no automatic solution to this end. When the human does this, we implement a system that sends a notification to *DeepFAQ* and then stores this new question answered. This action allows us to be updated with new products, pairs of question and answers by promoting continuously the system evolution, and thus, answering more questions.

4. Experimental Evaluation

Our system was evaluated based on real-world stores that sell various products, such as clothes, car parts, electronics, and furniture. The evaluation covered the entire solution, currently deployed on the largest e-commerce platform in South America. All stores are utilizing our DeepFAQ tool in their production process. We gathered quantitative results to understand how crucial DeepFAQ is for these stores. Our main focus is to provide reports on the effectiveness of responses and the cash conversion rate connected to customer inquiries. Essentially, we analyzed whether questions answered by deepFAQ result in product sales.

4.1. Overall effectiveness on a real scenario

This analysis aimed to understand how many questions our system could answer and how these results automatize the human attendance process. In this sense, the objective of this analysis was to measure the impact of the proposed solution, observing how well it does in conjunction with the Base System. The collected data corresponded from stores of 338 clients. All stores have their vector space databases, which could have more than 3 million questions. For this experiment, we evaluated the efficiency of DeepFAQ in actual stores by gathering data for one month. We chose this timeframe because our solution's application remained unchanged, with no significant software updates. In the period of our evaluation, DeepFAQ answered 13,991 questions.

Regarding its accuracy, since it would be very laborious to evaluate all these questions manually, we used the feedback from the stores themselves, which uses a review system to indicate if a response is correct or not. The revised question achieved a total

effectiveness of 97.75% from the questions answered by DeepFAQ during the evaluation time. Additionally, none of the stores using the system asked for it to be shut down in no time, indicating high user satisfaction.

Table 1 shows a summary of most common questions' intentions answered by DeepFAQ. These questions are related to specification, compatibility, availability, and others for intentions as greetings and acknowledge, which do not represent 1% of the total questions. We notice that the specification question means 41.33% of the question answered by DeepFAQ solutions. While compatibility, availability, and others categories represent 23.36%, 23.92%, and 11.39%, respectively. Moreover, DeepFAQ achieved a precision higher than 0.97 for each intention.

Table 1. Questions answered by DeepFAQ from May 14, 2023 to June 14, 2023

	Client feedback	
Intention	**n questions**	**Effectiveness (%)**
Overall	13991	97.75
Specification	5783	97.75
Compatibility	3269	98.01
Availability	3346	97.46
Others	1593	97.80

4.2. Cash conversion using DeepFAQ by countries marketplaces

In this analysis, we utilized a recently implemented report to demonstrate to our clients (marketplaces) how our system can help to sell their products. We are reporting on the Cash Conversion between May 14, 2023, and June 14, 2023. The instability of the dollar exchange rate in the long term has made it challenging to measure long past periods, which is why we have chosen to focus on a shorter period.

Table 2 presents the cash conversion of the questions managed by DeepFAQ in the four countries where DeepFAQ operates; we use a money exchange provided by Google on June 20, 2023 to convert the local currency to USD. We notice a significant conversion in Brazil regions where DeepFAQ helped marketplaces obtain **$45,177.07** while others regions got around 10% of this quantity; the total conversion amount is **$58,683.17**.

Table 2. Cash convertion using DeepFAQ from May 14, 2023 to June 14, 2023

	Cash conversion ($USD)
Argentina	4,939.07
Brazil	45,177.24
Chile	4,513.37
Mexico	4,053.49

5. Discussion

In our first analysis, we observed that *DeepFAQ* can perform well in a real-world scenario over a one-month period. Even though the number of answer question answered by DeepFAQ is "low" (13991), this can be explained by the absence of appropriate candidates for

a given question. A DeepFAQ solution relies on the existing appropriate candidates in the space to provide relevant answers. However, our solutions fix such problems for future questions applying the *Continuous populating* described in Section 3.

Our experiments showed that the system's answer effectiveness is **97.75%**, which reveals that our system can usually give a correct answer. Additionally, Table 1 indicated that client satisfaction remains high across questions with different intentions, demonstrating that our solution is not biased toward specific questions.

Marketplaces are interested in understanding how our solutions can assist them in resolving customer inquiries. Also, they want to know if our solutions generate sales for their stores, measuring their monetary value. Our solutions provide adequate answers to multiple intentions of questions, which makes them suitable for retrieving FAQs from the e-commerce domain. Our quality analysis indicated a high keyword sensibility that only allows answering a question with the most ranked candidate.

Our analysis demonstrated that the presence of DeepFAQ is helpful during the sales, being a crucial key factor because its absence can have noticeable effects, especially for Brazilian marketplaces in a few months. The low conversion rate in countries outside of Brazil may be attributed to the fact that the *GoBots* company initially began with Brazilian stores, which make up approximately 80% of all DeepFAQ-affiliated stores.

6. Conclusion

Question answering based on sentence similarity computation remains an open research challenge. This requires solutions to handle several domains simultaneously and adequately addresses informal issues raised in Portuguese language questions. This study presented and evaluated a system by combining semantic search with a neural network as a classification system to rank similarity scores. Our system retrieves similar questions based on vector space encoding. The proposal showed an effective solution to help to solve the problem of question-answering in e-commerce platforms with a small need for data labeling or large languages models services. Furthermore, working with existing software systems implemented in this context made it possible to significantly increase the number of questions answered by maintaining a good enough precision, considering the 97.75% of assertiveness in the answers given. Our evaluation process in a real-world setting showed that our solution is an up-and-coming technique to contribute to addressing question-answering in e-commerce platforms. In future work, we plan to overcome the problems of auto-populating Questions-Answers pair using Large Language Models. This might reduce the need to have human historical questions-answer pair and anticipate the need for inquiries from customers.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. In addition, this research was partially funded by the São Paulo Research Foundation (FAPESP) (grant #2022/13694-0).[3] We would like to thank GoBots for collecting data and sharing their environment.

[3]The opinions expressed in this work do not necessarily reflect those of the funding agencies.

References

Cer, D., Yang, Y., yi Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder.

Chen, S., Li, C., Ji, F., Zhou, W., and Chen, H. (2019). Review-driven answer generation for product-related questions in e-commerce. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 411–419, New York, NY, USA. Association for Computing Machinery.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Finardi, P., Viegas, J. D., Ferreira, G. T., Mansano, A. F., and Carid'a, V. F. (2021). Bertaú: Itaú bert for digital customer service. *ArXiv*, abs/2101.12015.

Fonseca, E. R., Borges dos Santos, L., Criscuolo, M., and Aluísio, S. M. (2016). Visão geral da avaliação de similaridade semântica e inferência textual. *Linguamática*, 8(2):3–13.

Gormley, C. and Tong, Z. (2015). *Elasticsearch: The Definitive Guide*. O'Reilly Media, Inc., 1st edition.

Gupta, M., Kulkarni, N., Chanda, R., Rayasam, A., and Lipton, Z. C. (2019). Amazonqa: A review-based question answering task. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4996–5002. International Joint Conferences on Artificial Intelligence Organization.

Gupta, S. and Carvalho, V. R. (2019). Faq retrieval using attentive matching. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 929–932, New York, NY, USA. Association for Computing Machinery.

Kulkarni, A., Mehta, K., Garg, S., Bansal, V., Rasiwasia, N., and Sengamedu, S. (2019). Productqna: Answering user questions on e-commerce product pages. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 354–360, New York, NY, USA. Association for Computing Machinery.

Mass, Y., Carmeli, B., Roitman, H., and Konopnicki, D. (2020). Unsupervised FAQ retrieval with question generation and BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 807–812, Online. Association for Computational Linguistics.

Sakata, W., Shibata, T., Tanaka, R., and Kurohashi, S. (2019). Faq retrieval using query-question similarity and bert-based query-answer relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1113–1116, New York, NY, USA. Association for Computing Machinery.

Shinzato, K., Shibata, T., Kawahara, D., Hashimoto, C., and Kurohashi, S. (2008). TSUBAKI: An open search engine infrastructure for developing new information access

methodology. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Hernandez Abrego, G., Yuan, S., Tar, C., Sung, Y.-h., Strope, B., and Kurzweil, R. (2020). Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.