Automatic Spanish Vocabulary Generation and Resource Recommendation for Nonnative Language Learners

Mohit Garg
CS 4650 Final Project
Georgia Institute of Technology

Abstract

The process of learning a new language can be challenging and time-consuming, especially when it comes to acquiring a rich vocabulary and finding suitable resources tailored to individual needs. In this paper, we propose a novel approach to automatically generate context-specific vocabulary lists and recommend relevant resources for Spanish language learners based on their input text. Our system leverages a fine-tuned BERT model for topic classification, extracting keywords from a preprocessed dataset of Spanish words and phrases, and subsequently identifying appropriate resources such as podcasts, articles, or videos. The results demonstrate the potential of our approach in providing personalized vocabulary lists and resource recommendations, facilitating a more effective and engaging language learning experience for users.

1 Introduction

The ability to communicate effectively in multiple languages has become increasingly important in today's globalized world. Learning a new language, however, can be a daunting task, particularly when it comes to building a strong vocabulary foundation and identifying relevant learning resources. Recent advances in natural language processing (NLP) and machine learning have paved the way for innovative approaches to language learning, offering new opportunities to enhance the learning experience for users.

In this paper, we present a context-driven system that automatically generates personalized vocabulary lists and recommends appropriate resources for Spanish language learners based on their input text. The primary motivation behind our approach is to provide a more targeted and engaging learning experience, allowing users to focus on vocabulary and resources that are relevant to their specific needs and interests.

Our system employs a fine-tuned BERT model for topic classification, which is used to predict the topic of the input text. This information is then leveraged to extract relevant keywords from a preprocessed dataset of Spanish words and phrases. Subsequently, the system identifies suitable resources, such as podcasts, articles, or videos, that align with the predicted topic and keywords, providing users with a comprehensive and context-driven learning experience.

2 Background and Related Work

Language learning has been a subject of interest for researchers and educators alike, with numerous methods and approaches being developed over the years to facilitate the process. With the advent of computational linguistics and natural language processing, a new set of opportunities has emerged to enhance and personalize language learning experiences. In this section, we provide an overview of the background and related work in the fields of language learning, NLP, and machine learning.

2.1 Traditional Language Learning Approaches

Traditional language learning approaches often rely on classroom instruction, textbooks, and language courses that follow a predetermined curriculum. While these methods have proven effective to some extent, they often lack the ability to adapt to individual learners' needs, interests, and learning styles. Furthermore, the process of acquiring a new language can become monotonous and unengaging, leading to reduced motivation and suboptimal learning outcomes.

2.2 Computer-Assisted Language Learning (CALL)

CALL encompasses the use of technology to assist in language learning processes. CALL systems and applications typically offer various tools and resources, such as vocabulary building exercises, grammar drills, and listening comprehension activities. While CALL has been successful in delivering personalized and engaging learning experiences to some extent, there is still room for improvement, especially when it comes to understanding and catering to the context in which learners are operating.

2.3 Natural Language Processing and Machine Learning in Language Learning

In recent years, NLP and machine learning techniques have been increasingly employed in language learning applications. This includes approaches such as intelligent tutoring systems, adaptive learning platforms, and personalized learning experiences. For instance, researchers have used NLP to automatically generate vocabulary exercises or to create personalized reading lists based on learners' interests and proficiency levels.

2.4 BERT and Transfer Learning

The introduction of BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2018) has revolutionized the field of NLP, setting new benchmarks in a wide range of NLP tasks. BERT's pre-training and fine-tuning process allows the model to be adapted to specific tasks with relatively small amounts of training data. This transfer learning approach has been used in various language learning applications, such as topic classification, sentiment analysis, and text generation.

3 Methods

In this section, we describe the methodology employed in our project to generate personalized vocabulary lists and resource recommendations for Spanish language learners. Our approach consists of two main components: topic classification using a fine-tuned BERT model and generating vocabulary lists and resource recommendations based on the identified topic.

3.1 Data Preprocessing and Preparation

The first step in our methodology involves preprocessing and preparing the dataset for training and evaluation. We start by cleaning the text data, which includes removing special characters, punctuations, and converting the text to lowercase. We then tokenize the text and create a set of unique keywords for each topic category.

3.2 Topic Classification with BERT

We use a pre-trained BERT model (dccuchile/bert-base-spanish-wwm-uncased) as the basis for our topic classification model. We fine-tune the model on our preprocessed dataset, which contains text samples and corresponding topic categories. The fine-tuning process involves training the model on our dataset, with the goal of adapting the pre-trained BERT model to the specific task of topic classification for our context.

3.3 Generating Personalized Vocabulary Lists

Once the topic classification model is trained, we use it to predict the topic category for a given input text provided by the user. Based on the identified topic, we generate a personalized vocabulary list by selecting a set of relevant words from our dataset that match the predicted topic. This ensures that the generated vocabulary list is tailored to the user's context and interests.

3.4 Generating Resource Recommendations

In addition to generating personalized vocabulary lists, we also provide users with resource recommendations that are relevant to the identified topic. These recommendations include online resources such as podcasts, articles, and videos that can help users improve their language skills in the context of the predicted topic. We curate these resources and associate them with the corresponding topic categories in our dataset. When a topic is identified for a given input text, we retrieve the resources associated with that topic and present them to the user as recommendations.

3.5 Evaluation Metrics

To evaluate the performance of our topic classification model, we use standard classification metrics such as accuracy, precision, recall, and F1-score. These metrics allow us to assess the model's ability to correctly identify the topic category for a given input text. In addition, we use qualitative evaluations to assess the relevance and usefulness of the generated vocabulary lists and resource recommendations. This includes soliciting feedback from users and language learning experts to ensure that our system provides valuable and context-driven learning experiences.

4 Results and Evaluation

In this section, we present the results of our topic prediction model, which was trained on a dataset of Spanish texts. We used a pre-trained BERT model fine-tuned for topic classification. The metrics we used for evaluation are accuracy, precision, recall, and F1-score. The results are presented in Table 1.

Metric	Value
Loss	6.5097
Accuracy	0.7764
Precision	0.5315
Recall	0.5764
F1-score	0.4689
Runtime	54.9897 seconds
Samples/sec	130.515
Steps/sec	16.33

Table 1: Evaluation metric

Our model achieved an accuracy of 77.64%, indicating a high rate of correct topic predictions. However, the precision and recall values are comparatively lower at 53.15% and 57.64% respectively. This suggests that the model might not be performing well in distinguishing between certain topics, leading to a lower F1-score of 46.89%. To further improve the model's performance, we could consider the following refinements and optimizations:

Data augmentation: Expanding our dataset by creating new examples through techniques such as paraphrasing, back-translation, or synonym replacement can improve the model's ability to generalize to unseen examples.

Hyperparameter tuning: Exploring different combinations of hyperparameters, such as learning rates, batch sizes, and optimizer settings, can lead to better model performance.

Using more advanced architectures: Experimenting with other pre-trained models, such as RoBERTa or XLM-R, might yield better results for this task.

Feature engineering: Extracting additional features from the input text, such as n-grams, part-of-speech tags, or sentiment scores, could help improve model performance by providing more informative inputs.

Error analysis: Investigating specific cases where the model struggles to make accurate predictions can provide insights into potential areas for improvement. This might involve examining confusion matrices or analyzing misclassified examples to identify common patterns or issues.

5 Discussion

In this study, we have developed a topic prediction model using a pre-trained BERT model fine-tuned for topic classification on Spanish texts. Our model demonstrates promising results in predicting topics, which can be used to generate vocabulary lists and recommend learning resources to language learners. However, there is still room for improvement, as evidenced by the model's lower precision, recall, and F1-score values.

The results of our study indicate that deep learning approaches, such as BERT, can be effectively applied to the task of topic prediction in the context of language learning. As mentioned in the Results and Evaluation section, there are several potential refinements and optimizations that could be explored to enhance the model's performance further.

One limitation of our approach is the reliance on a pre-processed dataset, which might not fully represent the diversity and nuances of real-world Spanish texts. Additionally, the dataset might contain biases that could affect the model's ability to generalize to new, unseen examples. Future work could explore using more diverse and extensive datasets to train the model, which might lead to improved performance.

Another area for future research is the integration of our topic prediction model with resource recommendation systems. By combining our model's predictions with a database of learning resources, we could develop a more comprehensive solution for language learners, offering not only vocabulary lists but also targeted recommendations for podcasts, videos, articles, and other relevant resources.

Furthermore, our model could be extended to other languages and topics, allowing for a more personalized and adaptive language learning experience. By incorporating user feedback and dynamically updating the model's predictions, we could create a truly adaptive system that continually evolves to meet the needs of individual learners.

6 Potential Applications

Personalized Learning Resources: Our model can be used by language learning platforms to curate personalized resources for learners. By understanding the learner's interests, the system can recommend relevant content, such as articles, videos, podcasts, and quizzes, to make the learning experience more engaging and effective.

Course Material Generation: Educators can utilize the model to organize and categorize content for their courses. By identifying relevant topics and resources, teachers can create tailored lesson plans and learning materials for their students, ultimately improving the quality of their instruction.

Language Exchange Community: Our model can be integrated into language exchange platforms that connect Spanish learners with native speakers. By identifying users' interests, the platform can match learners with partners who share similar interests, fostering a more enjoyable and meaningful exchange experience.

Content Curation for Language Learning Apps: Language learning applications can leverage our model to categorize and filter content based on topics. This allows learners to focus on specific themes or subjects they are interested in, making their language learning journey more targeted and engaging.

Assessment and Progress Tracking: By analyzing the texts and topics that learners engage with, our model can help educators and learning platforms assess learners' progress and adapt their learning plans accordingly. This can result in more efficient learning experiences and better overall outcomes.

7 References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Ghemawat, S. (2016). TensorFlow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

Canales, L., & Reyes, A. (2020). Spanish Pre-Training BERT. DCC UChile. Retrieved from https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. 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), 4171-4186.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Reimers, N., & Gurevych, I. (2017). Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 338-348.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Louf, R. (2020). Transformers: State-of-the-art natural language processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38-45.

8 Acknowledgements

I would like to express my sincere gratitude to my professor, Dr. Wei Xu, for her valuable guidance and support throughout the course of this project. Her insights and feedback were instrumental in shaping the direction of this work. I would also like to thank my friend, Octavio Calvo, for his invaluable assistance with Spanish interpretation and structure. His expertise in the language greatly contributed to the overall quality and accuracy of the research.