Large Language Models for Difficulty Estimation of Foreign Language Content with Application to Language Learning

No Author Given

No Institute Given

Abstract. We use large language models to aid learners enhance proficiency in a foreign language. We achieve this by discovering content on topics the user is interested in and that correspond to a learner's degree of knowledge of that language. Our work centers on French content, but our approach is readily transferable to other languages. Our solution offers several distinctive characteristics that differentiate it from existing language-learning solutions, such as, a) the discovery of content across topics that the learner cares about, thus increasing motivation, b) a more precise estimation of the linguistic difficulty of the content than traditional readability measures, and c) the availability of both textual and video-based content. The linguistic complexity of video content is derived from the video captions. It is our aspiration that such technology will enable learners to remain engaged in the language-learning process by continuously adapting the topics and the difficulty of the content to align with the learners' evolving interests and learning objectives.

A video showcasing our solution can be found at:
`https://youtu.be/O6krGN-LTGI`

Keywords: digital education · competency-based learning · extensive reading · machine learning · large language models

1 Introduction

The value of gaining communicative competence is widely acknowledged, especially in today's interconnected world. Whether it is for expanding one's social circle, finding a job, as a hobby, or even as a protective mechanism against cognitive decline [18], learning a foreign language is a rewarding but also challenging task. A major hurdle pertains to the time required to build up the new vocabulary and to use the novel linguistic patterns in a fluid manner. Another challenge arises from the fact that learning a new language via the traditional way of using textbooks is a non-personalized and often unengaging process. This is because textbook content is static and follows a linear structure, unable to adapt to the learner's changing interests, requirements, and proficiency in the foreign language.

Contemporary online foreign-language educational tools, such as Duolingo, Frantastique, ReadLang, and others, integrate elements of gamification into the

learning process to enhance their appeal and to ultimately decrease learner attrition from learning platforms [29].

However, existing tools still lack important personalization aspects because they either: work with predefined generic materials (Duolingo, Frantastique), or expect the user to search for their own textual content on the internet (ReadLang). In our solution, we present an approach to address the aforementioned shortcomings. We leverage Internet resources to discover both video and textual content that the reader is interested in. We use machine learning techniques to estimate a) The general topic(s) that the digital contents covers, b) The linguistic "difficulty" for each unit of content, in the sense of how proficient a person should be in that foreign language to understand the vocabulary of the content. Over time, the system will understand better both the user's interests (thus recommending topics the user cares about) and the level of knowledge of the foreign language and only recommend content that is at the learner's level or slightly more difficult than their knowledge level (e.g., between A1 and C2, according to CEFR).

From a pedagogical perspective, our solution builds upon the already established theory of *extensive reading*. A large body of research both by Bamford and Day, as well as by Krashen, has shown that extensive reading constitutes a crucial means of reinforcing one's language skills, not only improving reading and vocabulary skills, but also yielding more comprehensive improvements across all areas of language competence [8,19]. Alan Maley, a distinguished English scholar, has even advocated extensive reading as the "single most important way to improve language proficiency" [23]. From a technological perspective, our work makes the following **contributions**:

1. We provide a solution using modern machine learning techniques to estimate the **difficulty** and the **topic** of digital content. Our method is based on state-of-the-art methods in text embeddings and large language models to estimate the text difficulty. It also discovers content on topics that the reader is interested in.

2. Previous works primarily addressed books or articles, but our solution is more comprehensive in that it retrieves also available **video content** from YouTube, and uses automatic captioning to infer its linguistic difficulty. In this manner, our approach benefits from the broad availability of engaging content in video format.

2 Pedagogical Underpinnings

Our learning approach is founded on several pedagogical principles. The first one being extensive readings (also known as Free Voluntary Reading [19]) which encourages students to choose their own reading materials and read for pleasure. A list of key principles for successful extensive reading has been proposed by [8] and then extended by [27] which we reiterate in Figure 1.

The reading workshop model [1] is another approach that involves learners reading a lot on topics they like. This approach allows students to choose books

1. Students read a lot and read often.
2. There is a wide variety of text types and topics to choose from.
3. The texts are not just interesting: they are engaging/compelling.
4. Students choose what to read.
5. Reading purposes focus on: pleasure, information and general understanding.
6. Reading is its own reward.
7. There are no tests, no exercises, no questions and no dictionaries.
8. Materials are within the language competence of the students.
9. Reading is individual, and silent.
10. Speed is faster, not deliberate and slow.
11. The teacher explains the goals and procedures clearly, then monitors and guides the students.
12. The teacher is a role model. . . a reader, who participates along with the students.

Fig. 1. Principles of extensive reading

based on their own interests and then participate in book talks, book clubs, and other reading-based activities to encourage discussion and analysis of the texts they have read.

Project-based learning is an approach where learners are encouraged to select a topic they are interested in and read extensively on it as part of the project [3]. This approach emphasizes the development of skills such as critical thinking, problem-solving, and collaboration while allowing learners to pursue their own interests and passions through reading.

Competency-based learning [6,2] is an educational approach that focuses on the mastery of specific skills or competencies, rather than just completing a certain number of courses or hours of study. Students are empowered daily to make important decisions about their learning experiences, how they will create and apply knowledge, and how they will demonstrate their learning. Students receive timely, differentiated support based on their individual learning needs, and they learn actively using different pathways and varied pacing.

As we elaborate on our solution, it will become clear to the reader that our approach directly addresses the first ten principles of extensive reading and is drawing upon several tenets of the aforementioned pedagogical approaches: we discover content that the learner cares about, reading/studying is done as a pleasure, the content is differentiated per learner and is selected to be within the learner's language competence.

3 Our approach and technological underpinnings

We describe our methodology and the architecture of our solution.

Estimating Difficulty/Level of text: We model the estimation of difficulty as a classification problem. So, if $\mathcal{D}$ is the set of documents to be classified, and

let Y be the random variable representing the linguistic difficulty class. Y For example, if we label difficulty based on the CEFR levels, Y can take values from the set $\{A1, A2, B1, B2, C1, C2\}$. A classifier is a function

$$f : \mathcal{D} \rightarrow Y$$

that maps a document d to a linguistic difficulty class $y \in Y$. The classifier f is built using pairs of text and a label, where the label corresponds to the linguistic difficulty of the content. To create this model that predicts the difficulty of foreign content, we use modern transformer neural networks. For example, a modern transformer architecture such as BERT [9], will convert the tokens of a text sentence into a series of 768-dimensional vectors capturing the "meaning" of each token (see Figure 2). Use of the different transformer models will results into vectors, or embeddings, of different lengths. For example the "ada-002" embeddings produce vectors of length 1536 and consider a context of more than 8000 token, thus they can encode more accurately each word based on its surroundings, ie the topic and the context that it covers.

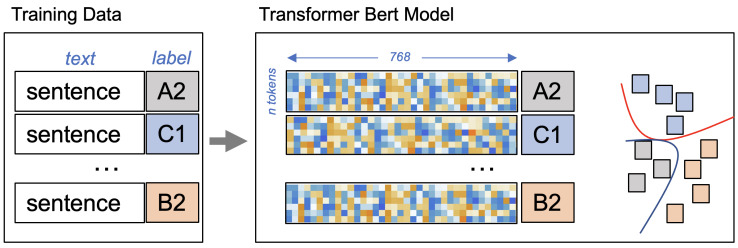

Fig. 2. Posing difficulty estimation as a classification problem.

Modern text embeddings based on transformer networks are *contextual* embeddings because they take into account the neighborhoods of words and subwords to determine the appropriate embedding values. They accomplish this by capitalizing on a self-attention mechanism, as illustrated in Figure 3. Earlier embedding techniques such as Glove [26] and Word2Vec [24] did not consider the context and thus create embeddings of inferior quality. As an example of a contextual embedding, if we have three sentences, two of them talking about Apple the company and one referring to the fruit, the texts that refer to the company will be mapped closer to each other, thanks to the disambiguation provided by the context.

Note, that the original embeddings are only used as seeds for the final model. The neural network model that is created will be post-trained more epochs using the new training labels provided. This process, also called "transfer-learning", leads to a more efficient training and creation of the new neural network because we do not need to start from scratch, but we tune further the model to adjust to the goals of our problem.

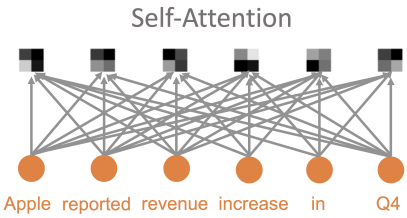

Fig. 3. Illustration of self-attention used by Bert embeddings

One shortcoming of these large language models, as their name suggests, is their large size. For example, models based on the BERT transformer model have sizes in the order of 500MB. One can further compress the model created using quantization and pruning techniques, a task also known as "knowledge-distillation" [12]. For example, there exist compressed versions of popular models, such as the TinyBERT [17] or DistilBert [28] for the BERT model. There is an increasing number of repositories that offer distilled version of already created models, one example being SparseZoo [1]. Such compressed models can be up to be 10 times smaller than the original models, and in practice lead to improvements in prediction accuracy, because of the inherent regularization offered via the compression process.

GPT-3+ Models: During the last years, even more advanced large-language models have been introduced in the Natural Language Processing (NLP) literature. Very prominent is the Generative Pre-trained Transformer 3 and 4 by OpenAI, or GPT-3 and GPT-4 for short [4], which have been trained on massive datasets of multilingual text to learn patterns and relationships in text and has led to the development of the popular ChatGPT. GPT-3 and GPT-4, have been pre-trained for next-token prediction and are post-trained using reinforcement learning from human feedback. GPT-3 and GPT-4 models achieve state-of-the art performance across a broad array of tasks [5]. Such models are particularly good at generating human-like text, but they can also be used as a basis for building advanced classification models, particularly with post-training (called fine-tuning) on data labelled for the task at hand. In our experiments, we use the GPT-3 model (and not the newer GPT-4) because the OpenAI platform allows for fine-tuning, at the current time point, only on the GPT-3 model, something that is essential for our specific task. In the experiments, we see that the GPT-3-based model builds the most accurate difficulty predictors given the labelled data provided.

Topic detection: Detecting the topics that a digital content (article, video, etc) touches upon is necessary for our solution, to make sure that we align the users' interests with the right content. Note, that much of the digital content either is readily available as text, or can be converted to text: e.g., for a video one can use the captions using speech-to-text technology, for music one can use the

[1] `https://sparsezoo.neuralmagic.com`

lyrics, etc. Topic detection can also modelled as a classification problem. That is, one can provide several examples of text and their labels with topics that the text covers. Here, we do not need to create our own model nor do we collect our own data as we did for the difficulty estimation, but we use already pretrained models based on zero-shot text classification [32]. Such models have been shown to be very effective, reaching accuracies that exceed 95%. They also recognize a broad spectrum of topics. Because such models are typically trained on English text, and we are working with multilingual content, we use a further step which is translating the content into English before providing it to the topic classifier.

Note, that in some cases, the digital content itself already offers a categorization into topics. For example, many YouTube videos come pre-tagged with topics which can offer a very fine-grained topic classification. If topics are present in the content, they are merged with the topics identified by the topic classifier.

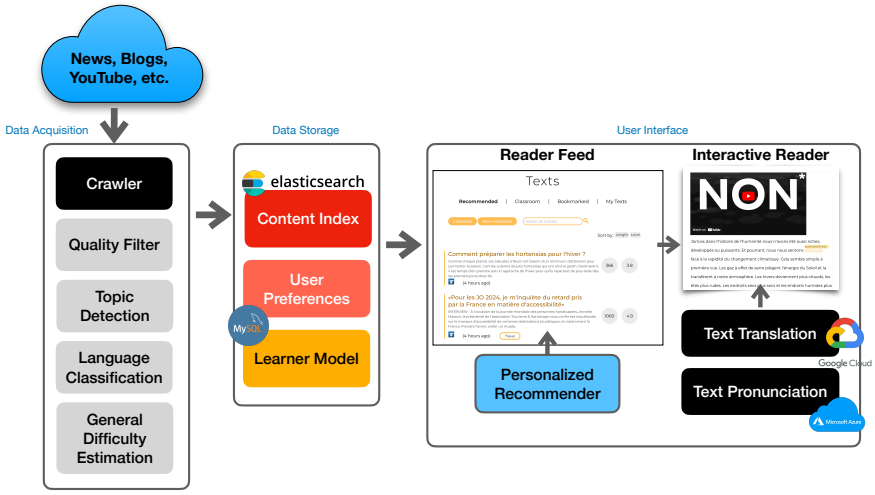

Fig. 4. Architecture of our approach. Data is acquired and classified into topic, language, and difficulty.

Table 1. Characteristics of the datasets in our experiments

Dataset	Total labels	Total words	Total characters	Labels
`LjL` [15]	2,060	334,026	1,532,442	level1,. . . , level4
`sentencesInternet`	4,800	85,941	421,045	A1, . . . , C2
`sentencesBooks`	2,400	56,557	281,463	A1, . . . , C2

3.1 Architecture

The general architecture of our solution is shown in Figure 4, consisting of data acquisition and cleaning components, data annotation components, data storage components, and user interface components. First, a **crawler** component monitors a series of RSS feeds (including the RSS of relevant YouTube channels) and every time a new item is discovered it passes it to a **quality filter**. The filter removes items that are behind a paywall, articles that have too little text, etc. Several other types of meta-information about the article are extracted by the crawler to be used later in the UI: authors, word count, and a short article description (usually the RSS feed item description is used for this).

The **topic detection** component extracts one or more topics from a text, both general and user defined. At the same time, the **general difficulty estimation** applies a series of difficulty estimation algorithms on the text, including the difficulty estimation as described in this work and also traditional readability measures [10].

The article together with the metadata, topics, and difficulty metrics are then added to the **text index** module which is implemented on top of an Elasticsearch cluster. From there, the **personalized recommender** creates a new mix of fresh recommendations every time the user logs into the web application. It takes into account the **user defined interests** and also the evolving **learner model** which tracks the current knowledge level of a particular user.

The **interactive reader** provides translation and pronunciation with the help of third-party **machine translation** services for words that the user needs and sends all these interactions together with explicit user feedback on text difficulty back to the **learner model** that updates the estimated knowledge level of the user.

3.2 User interface

We integrated the difficulty estimation algorithm presented in the previous sections on Zeeguu, an open source research project [22], which provides an easy to deploy platform for offering content to multiple users. Figure 5 presents the main components of the UI. The landing page of the application presents the recommended content for the learner, allowing the customization of the user interests. Users can declare both “interests“ and “non-interests“. In Figure 5, the user has queried (marked with ①) the phrase “intelligence artificielle” (i.e., Artificial Intelligence in French) and the system has returned a series of videos and articles. Upon opening an article, the user is sent to the original page of this article (③). If the learner also has the related reader extension installed on their browser, they can activate it (④) to remove the ads and the navigation from the page and focus only on the text which becomes interactive: words and phrases can be translated and pronounced. If, on the other hand, the user chooses to watch a video, the video is opened inside the web app (②) where the subtitles, are presented with interactive translation and pronunciation capabilities.

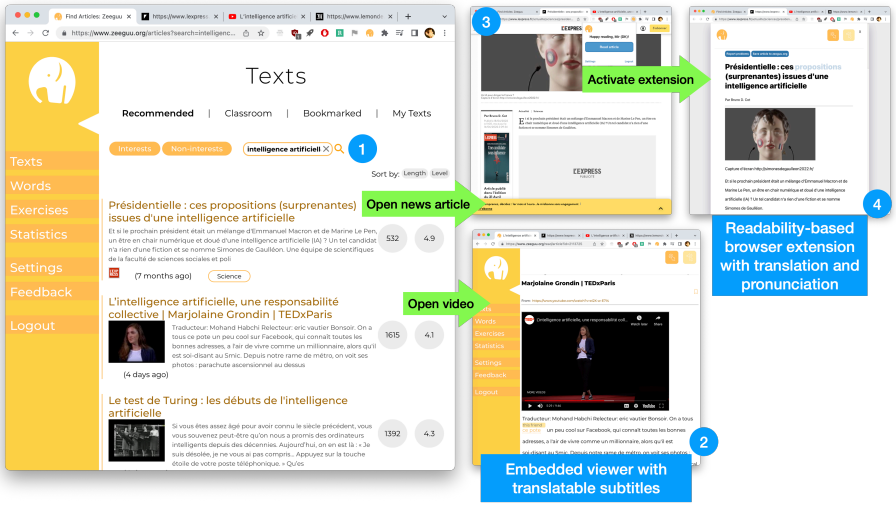

Fig. 5. The search results (1) summarize both news and video results. Videos can be watched in an embedded page (2) with translatable subtitles. When an article is being open, the reader is sent to the original source (3) where a Readability-like browser extension is available to clean up the page and support translation and pronunciation in the text (4)

4 Experiments

The goal of the following experiments is to demonstrate that the difficulty estimation technique presented can significantly improve the difficulty estimation offered by traditional readability metrics. We used three **datasets**:

1. Littérature de jeunesse libre (`LjL`) which we obtained from [15]. Each content item here contains several sentences and a label (labels: level1, level2, level3, level4).
2. A collection of sentences collected from the Internet (`sentencesInternet`). These were then annotated by at least three annotators. Only sentences in which all participants agreed on the difficulty annotation were retained (labels: A1,A2,B1,B2,C1,C2). Here, the labels correspond to the levels designed by the Common European Framework of Reference for Languages (CEFR) [2].
3. A collection of sentences from literature books (`sentencesBooks`). Each book was annotated with a difficulty level by a Professor of French. All sentences in that book were then given that label. This process involved an OCR pipeline which could lead to faulty detection of characters, so only the sentences without any errors were retained. (labels: A1,A2,B1,B2,C1,C2).

[2] `https://stay.fl-france.com/french-levels/`

The characteristics of these datasets are provided in Table 1. To train and evaluate our model, we use an 80/20 train-test split.

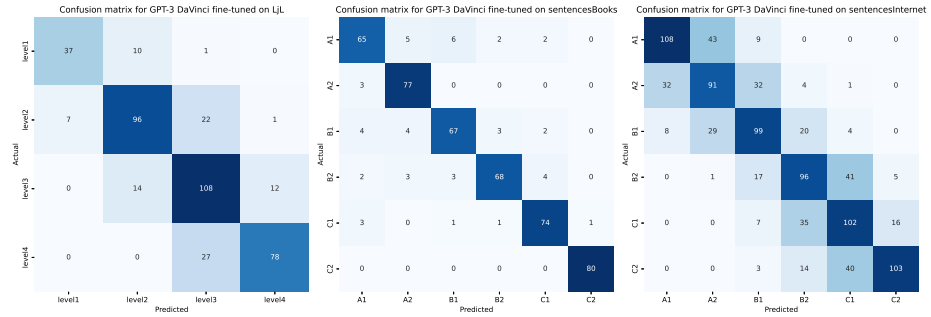

Fig. 6. The confusion matrix for the difficulty estimation technique based on the GPT-3-based difficulty-aware embeddings for the three datasets in our experiments.

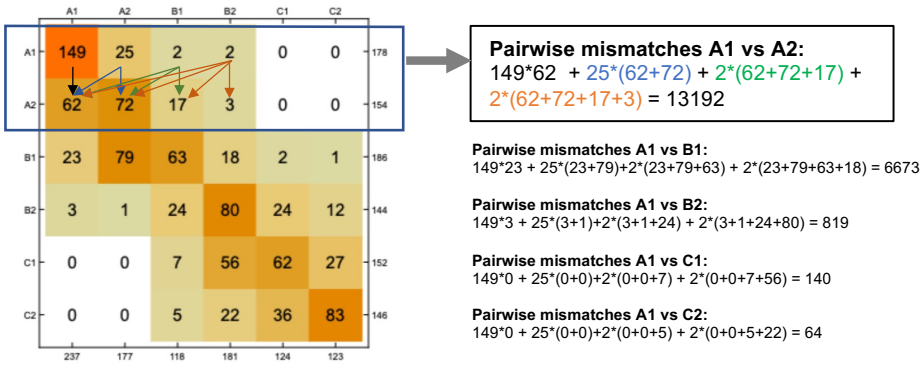

Fig. 7. Evaluating the pairwise mismatches for a given confusion matrix of a classifier.

4.1 Accuracy

We compare the accuracy of our difficulty estimation approach to traditional readability-based metrics, such as the GFI (Gunning Fog Index), FKGL (Flesch Kincaid grade level), ARI (Automated Readability Index). These techniques are inherently regression techniques and output a floating point value of the text difficulty, therefore we cannot make a direct comparison, because our difficulty estimator predicts discrete labels.

To make a meaningful quality comparison, we devise the following experiment which draws on **pairwise-comparisons**. For each text, we predict its

difficulty and compare it to all the other objects with different pre-annotated difficulty labels and record if this comparison was correct. For example, if a text with label A1 received an ARI score of 15 and a different sentence with (pre-annotated) label B2 received an ARI score of 14.6 (i.e., indicated as easier), then this comparison is considered incorrect and recorded accordingly. We call this incorrect comparison a **mismatch**. With our approach, we can also do the same by comparing the labels. If the confusion matrix of a classifier is already computed, then the pairwise mismatches can be evaluated as shown in Figure 7.

The results of this analysis are shown in Table 2. Higher numbers indicate more mismatches. Therefore, we see that compared to older or traditional methods of difficulty estimation, the approaches based on modern transformer neural networks offer a significantly lower error rate. For the GPT-3-based models, there exist several variants. Davinci is the most powerful one, while Curie tries to balance accuracy and speed. Both GPT-3 models have been fine-tuned for the task at hand using the examples with the labelled difficulties of the text. The GPT-3-based model offers a better estimation of difficulty compared to the BERT-based neural network model.

Table 2. Number of pairwise mismatches for the neural network difficulty estimation techniques compared to predominant readability metrics. Numbers in bold indicate fewer mismatches, and thus better estimation of difficulty.

	LjL	sentencesBooks	sentencesInternet
GFI	17,338	30,836	83,770
ARI	17,749	32,993	92,140
FKGL	19,634	31,475	85,306
BERT embeddings	11,354	9,875	67,288
GPT-3 Curie fine-tuned	11,635	7,055	60,165
GPT-3 DaVinci fine-tuned	**10,740**	**6,817**	**57,258**

We also provide a comparison of the overall prediction accuracy for the transformer based models in Table 4. The confusion matrix of the best performing GPT-3 DaVinci model across all three datasets is shown in Figure 6. The confusion matrix shows how the test examples were misclassified for each class. We observe that the majority of misclassifications happen for classes adjacent to the real class, that is, if the real difficulty of a text is C1, it is more likely that it is misclassified as B2 or C2 rather than as one of the other classes.

Table 3. Accuracy comparison across the difficulty-aware embeddings

	LjL	sentencesBooks	sentencesInternet
BERT embeddings	**0.79**	0.86	0.53
GPT-3 Curie fine-tuned	0.76	0.89	0.61
GPT-3 DaVinci fine-tuned	0.77	**0.90**	**0.62**

Table 4. Compression of difficulty-aware embeddings

Size		LjL	sentencesBooks	sentencesInternet
BERT embeddings [Devlin et al, 2018]	470MB **0.707**	0.747	0.493	
GPT-3 Curie fine-tuned	0.76	0.89	0.61	
GPT-3 DaVinci fine-tuned	0.77	**0.90**	**0.62**	

5 Related work

Our work relates to two topics: difficulty estimation of foreign language texts, and recommender systems. The aim of the first research topic is to develop tools that can help language learners and teachers select appropriate reading materials that match their proficiency level. The development of tools for estimating the difficulty of foreign language texts can help learners and teachers select appropriate materials that are challenging yet manageable, and can support the development of language proficiency over time.

One approach to estimating the difficulty of a foreign language text is to use readability formulas, which are mathematical algorithms that calculate the complexity of a text based on features such as sentence length and word frequency. Some commonly used readability formulas for foreign language texts include the Flesch-Kincaid Grade Level, the Simple Measure of Gobbledygook (SMOG), and the Gunning Fog Index. These technologies were initially developed for the English language, and gradually have been extended to other languages such as French, Chinese and Italian [25,7]. Note, that these readability formulas are primarily targeted to estimate the difficulty of a text for native speakers rather than for second language learners [31].

Another approach to difficulty estimation is to use machine learning techniques to analyze the text and predict its difficulty level based on various linguistic features. For example, researchers have used features such as syntactic complexity, word frequency, and semantic similarity to develop models that can accurately estimate the difficulty of foreign language texts [7,11]. A particularly notable advancement in this field in recent years is the integration of pre-trained word- and sentence- embeddings into text readability architectures [30,16,20].

In parallel, there has also been research on the recommendation of text documents based on the reader's interest and past behavior [13]. While earlier work in this domain applied techniques such as clustering and collaborative filtering [14], more recent efforts have moved towards the use of text embedding and reinforcement learning [21,33]. In our proposed solution, we combine these two lines of research to recommend interest-aware content at the appropriate level of difficulty based on the reader's comprehension skills.

6 Conclusion and next steps

Several existent pedagogical approaches advocate the freedom of learners to follow their own, personalized and varied learning pathway, reading on topics they are passionate about. The reason for this approach is to make sure that reading in a foreign language becomes a daily habit and is not done as a chore. We presented our solution of an educational recommendation engine that helps learners discover appropriate content in a foreign language and presents content that is on par with their knowledge of the foreign language. Our solution crawls Internet content (text and video), and matches content with users based on their preferences and their knowledge of the foreign language. Our experiments demonstrated that the difficulty estimation technique presented is significantly more accurate than existing and widely-used readability metrics. On a technological aspect, our work is the first one to use GPT-3-based generative neural networks to create difficulty-aware embeddings for foreign language text.

Our next step for this work is to conduct a large-scale evaluation of the tool at French-speaking universities in Europe. Our aspiration is that with an extensive deployment, we can assess the effectiveness of machine learning solutions in accelerating the learning process of not only foreign languages but also of general cognitive skills.

References

1. Atwell, N.: In the middle: A lifetime of learning about writing, reading, and adolescents. Heinemann (2015)
2. Book, P.A.: All hands on deck: Ten lessons from early adopters of competency-based education. Western Interstate Commission for Higher Education (2014)
3. Boss, S., Krauss, J.: Reinventing project-based learning: Your field guide to real-world projects in the digital age. International Society for Technology in Education (2022)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
5. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T., Zhang, Y.: Sparks of Artificial General Intelligence: Early experiments with GPT-4 (Mar 2023). https://doi.org/10.48550/arXiv.2303.12712, http://arxiv.org/abs/2303.12712, arXiv:2303.12712 [cs]
6. Burnette, D.M.: The renewal of competency-based education: A review of the literature. The Journal of Continuing Higher Education **64**(2), 84–93 (2016)

7. Chen, X., Meurers, D.: Ctap: A web-based tool supporting automatic complexity analysis. In: Proceedings of the workshop on computational linguistics for linguistic complexity (CL4LC). pp. 113–119 (2016)
8. Day, R., Bamford, J.: Top ten principles for teaching extensive reading (2002)
9. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1423, https://doi.org/10.18653/v1/n19-1423
10. François, T.: An analysis of a french as a foreign language corpus for readability assessment. In: Proceedings of the third workshop on NLP for computer-assisted language learning. pp. 13–32 (2014)
11. François, T., Fairon, C.: An "ai readability" formula for french as a foreign language. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. pp. 466–477 (2012)
12. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**, 1789–1819 (2021)
13. Guan, Z., Wang, C., Bu, J., Chen, C., Yang, K., Cai, D., He, X.: Document recommendation in social tagging services. In: Proceedings of the 19th international conference on World wide web. pp. 391–400 (2010)
14. Habibi, M., Popescu-Belis, A.: Keyword extraction and clustering for document recommendation in conversations. IEEE/ACM Transactions on audio, speech, and language processing **23**(4), 746–759 (2015)
15. Hernandez, N., Oulbaz, N., Faine, T.: Open corpora and toolkit for assessing text readability in french. In: Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference. pp. 54–61 (2022)
16. Imperial, J.M.: Bert embeddings for automatic readability assessment. arXiv preprint arXiv:2106.07935 (2021)
17. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: TinyBERT: Distilling BERT for natural language understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 4163–4174. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.findings-emnlp.372, https://aclanthology.org/2020.findings-emnlp.372
18. Klimova, B.: Learning a foreign language: A review on recent findings about its effect on the enhancement of cognitive functions among healthy older individuals. Frontiers in Human Neuroscience **12** (2018)
19. Krashen, S.D.: The power of reading: Insights from the research: Insights from the research. ABC-CLIO (2004)
20. Lee, J.S.: An editable learner model for text recommendation for language learning. ReCALL **34**(1), 51–65 (2022)
21. Liu, J., Dolan, P., Pedersen, E.R.: Personalized news recommendation based on click behavior. In: Proceedings of the 15th international conference on Intelligent user interfaces. pp. 31–40 (2010)
22. Lungu, M.F., van den Brand, L., Chirtoaca, D., Avagyan, M.: As we may study: Towards the web as a personalized language textbook. In: Proc. of CHI. p. 338 (2018)

23. Maley, A.: Extensive reading activities for the second language classroom. ELT Journal, 59(4) pp. 354 – 355 (2005)
24. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
25. Okinina, N., Frey, J.C., Weiss, Z.: Ctap for italian: Integrating components for the analysis of italian into a multilingual linguistic complexity analysis tool. In: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020. pp. 7123–7131 (2020)
26. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
27. Prowse, P.: Top ten principles for teaching extensive reading: A response. University of Hawaii National Foreign Language Resource Center (2002)
28. Shen, H., Zafrir, O., Dong, B., Meng, H., Ye, X., Wang, Z., Ding, Y., Chang, H., Boudoukh, G., Wasserblat, M.: Fast distilbert on cpus. arXiv preprint arXiv:2211.07715 (2022)
29. Tuncay, H.O.: App Attrition in Computer-Assisted Language Learning: Focus on Duolingo. Ph.D. thesis, McGill University (Canada) (2020)
30. Wilkens, R., Alfter, D., Wang, X., Pintard, A., Tack, A., Yancey, K.P., François, T.: Fabra: French aggregator-based readability assessment toolkit. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 1217–1233 (2022)
31. Xia, M., Kochmar, E., Briscoe, T.: Text readability assessment for second language learners. arXiv preprint arXiv:1906.07580 (2019)
32. Yin, W., Hay, J., Roth, D.: Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. arXiv preprint arXiv:1909.00161 (2019)
33. Zheng, G., Zhang, F., Zheng, Z., Xiang, Y., Yuan, N.J., Xie, X., Li, Z.: Drn: A deep reinforcement learning framework for news recommendation. In: Proceedings of the 2018 world wide web conference. pp. 167–176 (2018)