Evaluation and simplification of text difficulty using LLMs in the context of recommending texts in French to facilitate language learning

Henri Jamet
Faculty of Business and Economics
University of Lausanne, Switzerland
henri.jamet@unil.ch

Yash Raj Shrestha
Faculty of Business and Economics
University of Lausanne, Switzerland
yashraj.shrestha@unil.ch

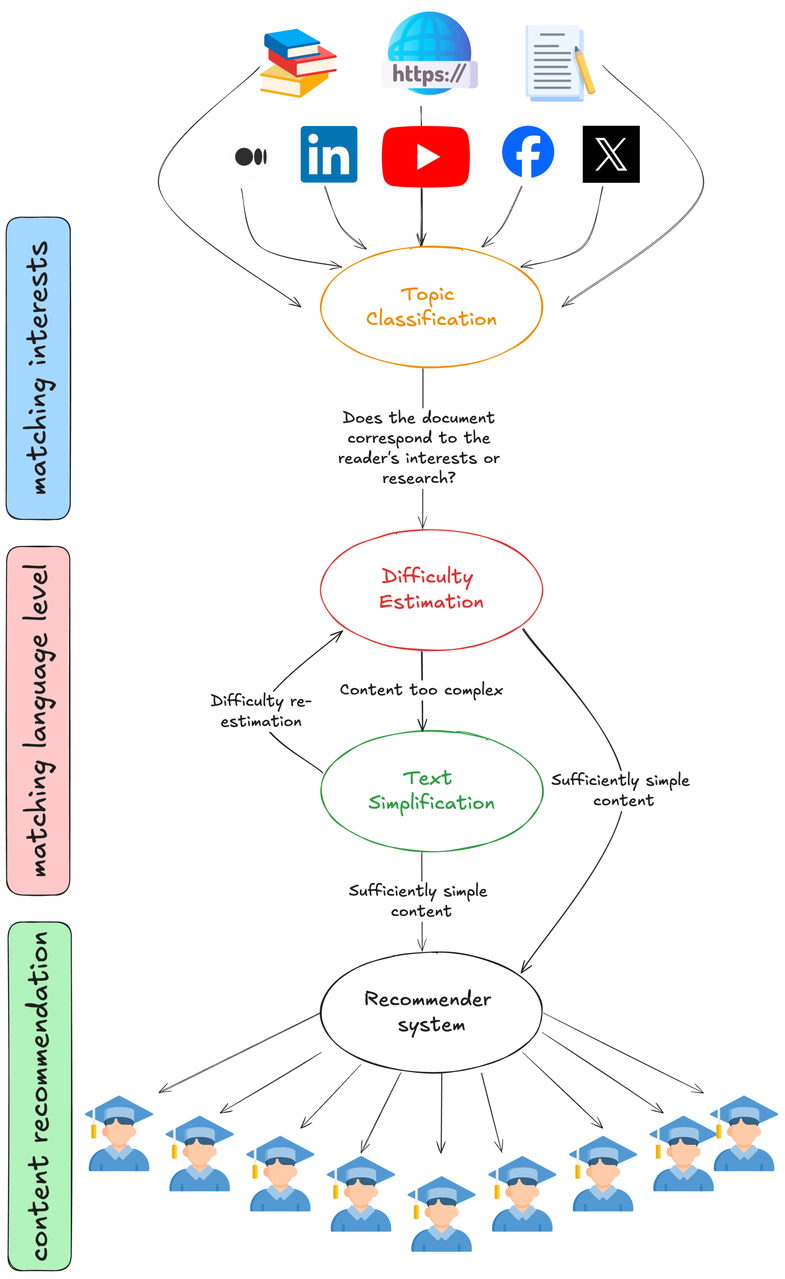
Get in touch with us

Maxime Manderlier
Faculty of Engineering
University of Mons (UMONS), Belgium
maxime.manderlier@umons.ac.be

Michalis Vlachos
Faculty of Business and Economics
University of Lausanne, Switzerland
michalis.vlachos@unil.ch

Paper and video

UNIL | Université de Lausanne
HEC Lausanne
UMONS
POLYTECH MONS



System Overview

Content Gathering: Texts in the target language are collected from various sources, including the internet, providing a broad range of material for recommendation.

Topic Filtering: Using Large Language Models (LLMs), we automatically classify the topics of these texts, filtering them based on the user's declared interests to ensure that the content aligns with their preferences.

Difficulty Estimation: We estimate the linguistic difficulty of each text using our CEFR-level prediction model. Only the texts that match the learner's current proficiency level are considered for recommendation.

Text Simplification: For texts that are too difficult but relevant to the user's interests, we apply an automatic text simplification model, which reduces the linguistic complexity while preserving meaning, expanding the pool of available content.

Final Recommendation: After filtering for both topic and difficulty, and simplifying as needed, the system recommends a curated set of texts tailored to the user's interests and language learning goals.

Introduction

We present a comprehensive approach to building a content recommendation platform designed to assist language learners. Our system is composed of four key components that work together to select, filter, and adjust texts to match the user's interests and language proficiency, ultimately facilitating personalized language learning.

Conclusion

By integrating these components, we provide a powerful recommendation system that ensures learners receive engaging and appropriately challenging materials, supporting their language acquisition journey with personalized, high-quality content.

Recommender system

Our system combines content-based and collaborative filtering to deliver personalized text recommendations. Using LLMs like BERT and ADA, we generate rich embeddings for texts and users, improving the alignment between content and user preferences. By integrating these embeddings into a graph-based model (LightGCN), we enhance the relevance and accuracy of recommendations, ensuring they adapt to the learner's evolving language skills and interests.

Text Simplification

We simplified texts to align with learners' proficiency levels while preserving meaning. Treating simplification as a sequence-to-sequence task, we fine-tuned an LLM using sentence pairs where each simplified sentence is one CEFR level lower than the original. With just 125 sentence pairs for fine-tuning, we significantly outperformed zero-shot LLMs. To evaluate quality, we introduced two metrics: simplification accuracy (effective difficulty reduction) and semantic similarity (meaning retention via cosine similarity), combined into a weighted score. Our experiments show that fine-tuned LLMs effectively simplify texts, expanding suitable content for learners at various proficiency levels.

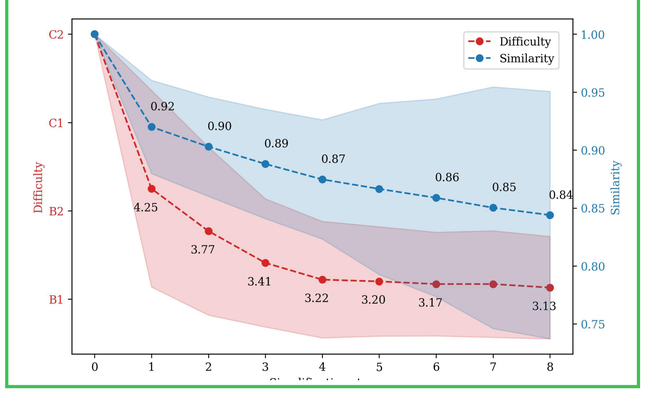
Difficulty Estimation

We framed automatic text difficulty estimation as a classification task to predict the CEFR level of a text. Using three annotated datasets, we compared traditional readability metrics adapted to French (FKGL, GFI, ARI) with logistic regression for class predictions, and tested various LLMs, including those trained on French data. The results show that LLMs significantly outperform traditional readability indices, which is crucial for recommending texts tailored to learners' proficiency levels.

model	context	LjL	SentencesInternet	SentencesBooks
GPT-3.5[1]	✓	0.72	**0.90**	0.50
	-	**0.73**	0.87	0.49
BERT[1]	-	0.62	0.82	**0.52**
Mistral[1]	✓	0.64	0.75	0.51
Davinci[1]	-	0.59	0.82	0.47
	✓	0.61	0.81	0.47
Mistral[1]	-	0.47	0.63	0.35
FKGL	-	0.42	0.34	0.35
GFI	-	0.45	0.32	0.34
ARI	-	0.40	0.34	0.34

Dataset	Model	Recall@5	Precision@5	F1@5	NDCG@5	MRR@5	MAP@5
Zeeguu	ALS	0.0626	0.0236	0.0309	0.0595	0.0774	0.0498
	BPR	0.0182	0.0061	0.0082	0.0168	0.0191	0.0146
	LMF	0.0251	0.0084	0.0116	0.0235	0.0291	0.0202
	LightGCN ADA (9 layers)	**0.0721**	**0.0274**	**0.0352**	**0.0666**	**0.0843**	**0.0550**
	LightGCN Xavier ADA (10 layers)	0.0630	0.0238	0.0308	0.0594	0.0759	0.0498
	LightGCN Bert (10 layers)	0.0659	0.0249	0.0323	0.0609	0.0788	0.0501
	LightGCN Xavier Bert (10 layers)	0.0622	0.0231	0.0300	0.0599	0.0765	0.0508
ml-100k	ALS	0.0767	0.1285	0.0806	0.1414	0.2504	0.0850
	BPR	0.0551	0.0975	0.0580	0.1084	0.1995	0.0634
	LMF	0.0381	0.0635	0.0400	0.0704	0.1419	0.0379
	LightGCN ADA (2 layers)	**0.0903**	**0.1348**	**0.0882**	**0.1556**	0.2682	**0.0975**
	LightGCN Xavier ADA (2 layers)	0.0881	0.1346	0.0869	0.1548	0.2720	0.0958
	LightGCN Bert (2 layers)	0.0855	0.1316	0.0847	0.1547	**0.2761**	0.0971
	LightGCN Xavier Bert (1 layer)	0.0855	0.1323	0.0847	0.1508	0.2636	0.0922
Goodreads	ALS	0.0634	0.0156	0.0237	0.0447	0.0419	0.0366
	BPR	0.0447	0.0114	0.0168	0.0323	0.0323	0.0261
	LMF	0.0432	0.0099	0.0155	0.0280	0.0245	0.0221
	LightGCN ADA (2 layers)	**0.0732**	**0.0183**	**0.0275**	**0.0513**	**0.0481**	**0.0415**
	LightGCN Xavier ADA (3 layers)	0.0603	0.0155	0.0231	0.0432	0.0415	0.0352
	LightGCN BERT (5 layers)	0.0673	0.0170	0.0254	0.0473	0.0449	0.0382
	LightGCN Xavier BERT (2 layers)	0.0570	0.0149	0.0220	0.0416	0.0405	0.0340
Tomplay	ALS	0.0875	0.0598	0.0650	0.0838	0.1408	0.0528
	BPR	0.0295	0.0236	0.0246	0.0306	0.0583	0.0178
	LMF	0.0327	0.0249	0.0263	0.0323	0.0598	0.0187
	LightGCN ADA (2 layers)	0.0958	0.0646	0.0706	0.0924	0.1541	0.0595
	LightGCN Xavier ADA (3 layers)	0.0943	0.0643	0.0699	0.0910	0.1528	0.0581
	LightGCN Bert (2 layers)	**0.0984**	**0.0660**	**0.0722**	**0.0947**	**0.1582**	**0.0608**
	LightGCN Xavier Bert (2 layers)	0.0966	0.0656	0.0715	0.0936	0.1566	0.0601

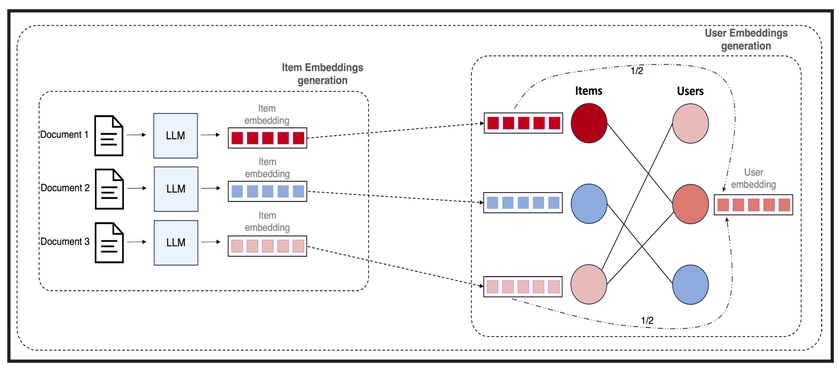
model	fine-tuned	Simplification Accuracy	Semantic Similarity	Weighted-Score
Mistral[2]	✓	**0.59**	0.91	**0.72**
GPT-3.5[2]	✓	0.57	0.91	0.70
GPT3.5[2]	-	0.53	0.93	0.67
GPT-4	-	0.51	**0.93**	0.66
Mistral[2]	-	0.47	0.93	0.63
GPT-4o	-	0.46	0.89	0.60
Davinci[2]	✓	0.44	0.83	0.57



Topic Classification

To ensure recommended content matches learners' interests, we developed an automatic topic classification method using LLMs to predict the main topic of a text. Utilizing a dataset of 1,743 text-label pairs from a language learning platform covering 11 categories (World, Travel, Music, etc.), we explored various models, including zero-shot and fine-tuned approaches. By assessing models of different sizes and architectures, we found that smaller, fine-tuned models specialized in French outperform larger, general-purpose ones in topic prediction accuracy. This enables our system to effectively filter and suggest content aligned with learners' declared interests.

model	accuracy
Flaubert-fine-tuned	**0.74**
GPT-4-turbo-2024-04-09	0.61
GPT-4o-2024-05-13	0.61
GPT-3.5-turbo-1106	0.58
Flaubert-pretrained	0.56
mDeBERTa	0.45
Davinci-002	0.09



The ACM Conference Series on Recommender Systems