idioma

Document-Based Learning in Second-Language Acquisition

A Thesis Defense by Paulo Frazão ‘20
Advised by Xiaoyan Li

Motivation and Goal

Cutting-edge technologies in an under-represented field

Motivation: A Different Kind of Learning

- High demand for reading comprehension development
  - Corr. with strong decoding & listening comprehension (Hagtvet 2003)
  - Especially true in learning a new language
    - Knowledge of target language key to reading comprehension ability (Verhoeven 2011)



- High demand for motivation-driven learning
  - McNamara (2012): “motivation should be included explicitly in instruction…[in] reading comprehension”
    - Corroborated by Logan (2011)
- Can come in many forms:
  - Gamification
  - Customization
  - Monetary rewards
  - Etc.

Goal: To develop a language-learning system that...

Motivates & excites students

Employs techniques like gamification and customization to ensure students keep pursuing their learning

Focuses on reading comprehension

Makes the development of this skill the primary objective, as opposed to a pedagogical afterthought

Leverages cutting-edge tech

Utilizes state-of-the-art libraries for machine learning, NLP, and other critical functions

Background & Related Work

Existing literature at three levels of granularity

Language-Learning Applications

Memrise

A flashcard-driven course repository that is extensive yet limited by its user-created content and form factor

Duolingo

A popular service with well-crafted courses and strong gamification hooks, which lacks only in its reading comprehension strategies

Reading Comprehension Systems

Lingua.com

A dedicated reading comprehension library with many texts but suboptimal customization and feedback options

Readlang

A browser extension allowing users to translate foreign texts on-the-fly, filling a unique, yet insufficient in the language market

Machine Learning and NLP in Readability

This project leveraged a number of papers in the surprisingly dense field of NLP for Portuguese text readability.

Automatic Text Difficulty Classifier
Assisting the Selection Of Adequate Reading Materials For European Portuguese Teaching

Pedro Curto[1,2], Nuno Mamede[1,2] and Jorge Baptista[1,3]
[1] *Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal*
[2] *INESC-ID Lisboa/L2F – Spoken Language Lab, R. Alves Redol, 9, 1000-029 Lisboa, Portugal*
[3] *Universidade do Algarve/FCHS and CECL, Campus de Gambelas, 8005-139 Faro, Portugal*
{pedro.curto, nuno.mamede}@l2f.inesc-id.pt, jbaptis@ualg.pt

Automatic Construction of Large Readability Corpora

Jorge Alberto Wagner Filho, Rodrigo Wilkens and Aline Villavicencio
Institute of Informatics, Federal University of Rio Grande do Sul
Av. Bento Gonçalves, 9500, 91501-970, Porto Alegre, RS, Brazil
{jawfilho,rodrigo.wilkens,avillavicencio}@inf.ufrgs.br

Approach

Bringing reading comprehension, gamification, and machine learning together

Idioma: A Web App for Reading Comprehension

Central Conceit: A web application that leverages NLP classifiers and a proprietary algorithm to offer a student documents tailored to their experience and interests

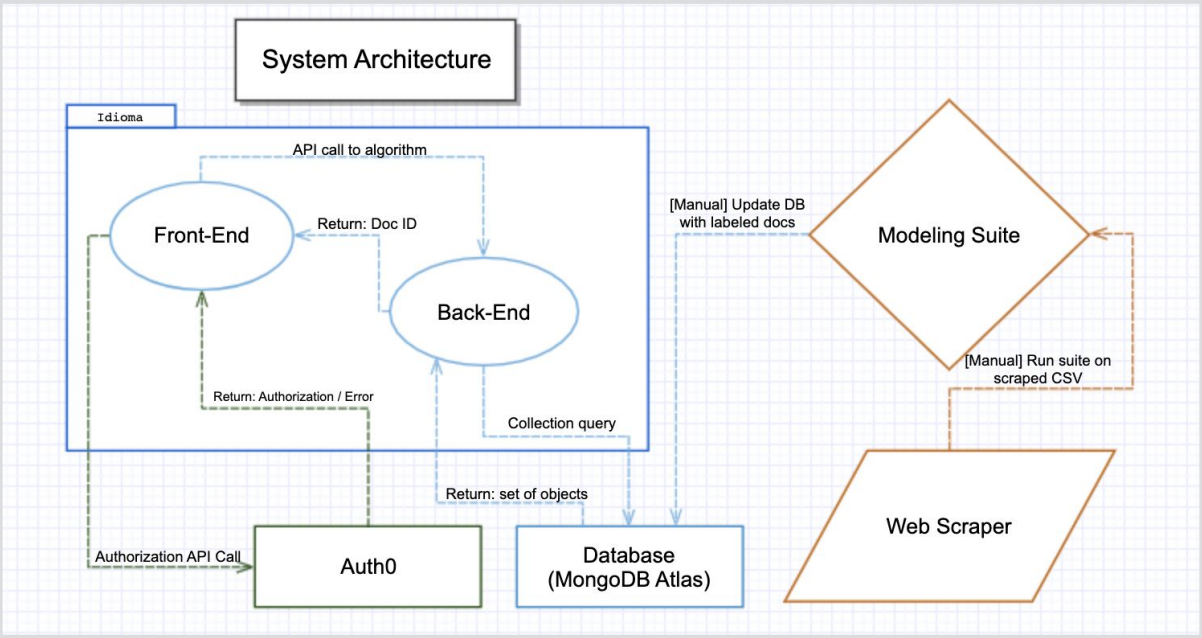
Advantages:

1. Centralizes reading comprehension
2. Enables use of leading tech, especially in ML and NLP
3. Foregrounds user agency & retention through gamification and customization
4. Presented in an intuitive and accessible web interface

Implementation

The inner workings of the Idioma application

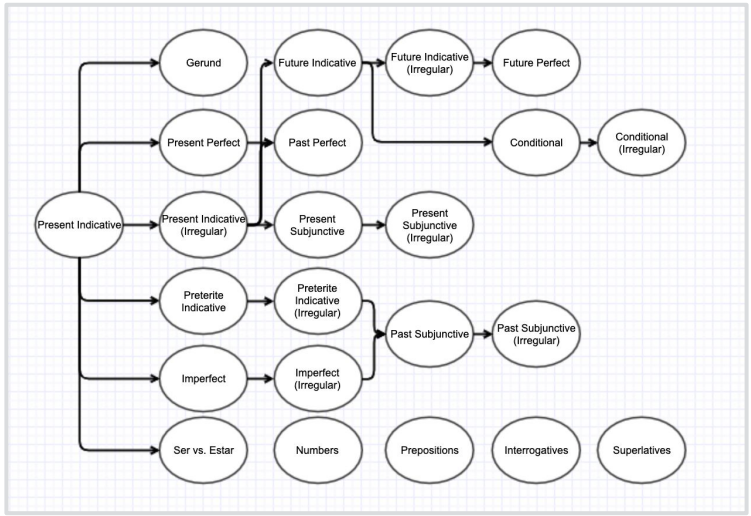
Construction Overview



Article Selection Algorithm

Pipeline:

- “Relevance” of each grammatical feature at right is calculated
- 100 articles sampled from set
- “Prevalence” of each feature at right in each article is calculated
- “Relevance” and “Prevalence” aggregated into a normalized “Appropriateness” score
- Highest-scoring article returned to user



Demo!

Let us step through a student's potential first session on the Idioma application

Data, Models & Results

The data, the models, and the results that they enabled

Data & Modeling Suite

Training Data:

- Supplied by Jorge Filho
- ~5,300 instances (documents)
- Labeled ‘difficult’ or ‘simple’

Instances (like example pictured below) fed to the modeling suite classifiers during training

Implemented Classifiers:

- Decision Tree
- Random Forest
- Bagging
- Gaussian Naive-Bayes
- K-Nearest Neighbor
- Multi-layer Perceptron
- Support Vector Classifier
- SGD Classifier

words	syllables	letters	unique	ttr	flesch-adap	coleman	flesch-grade
765	1539	3781		0.4339869281	54.29075066	11.90755556	16.89686017

ari	awl	awlstd	psfl	zh	wiki	brescola
12.77766947	4.94248366	3.164234211	d	d	s	s

Model Performance

Chosen Implementation:
Best-of-3 approach using Bagging, Random Forest, and MLP

Notable Observations:
- High, yet inverse, precision-recall disparities in SGD and Gaussian NB
- Bagging performs very well using very few features

Metrics and Statistics	% Features Used	Accuracy	Precision	Recall	F-measure
Decision Tree	0.6	0.73	0.68	0.69	0.68
Stochastic Gradient Descent	0.65	0.55	0.49	0.98	0.65
Multi-Layer Perceptron ☆	0.5	0.81	0.72	0.88	0.79
Gaussian (Naive Bayes)	0.5	0.7	0.8	0.35	0.48
Bagging ☆	0.35	0.79	0.74	0.74	0.74
Random Forest ☆	0.55	0.85	0.79	0.84	0.81
Support Vector Classifier	0.45	0.79	0.71	0.86	0.78
k-Nearest Neighbor Classifier	0.65	0.8	0.74	0.81	0.77

Conclusion & Future Work

Looking backward at our development and forward towards future extensions

Self-Assessment

Idioma's current iteration meets & exceeds all of the project objectives:

- Motivates learners through achievements & greater agency over their learning path
- Lays foundation for consistent, challenging reading comprehension practice
- Leverages cutting-edge tech, and can easily be extended to keep doing so in the future

Future Work

Opportunities for future contribution:

- Extension of web scraper to other hosts
- Expansion of modeling suite to keep pace with ML innovation
- Refinement of authentication system
- Construction of more practical and useful data set
- Algorithmic and parametric tweaks

Code and data set available at:
https://github.com/paulo892/IdiomaFinal


Thanks for listening!
Any questions?
idioma

Credits - SlidesGo.com

See below for credits for the template used for this presentation:

Presentation template by Slidesgo

Icons by Flaticon

Images & infographics by Freepik

Big image slide photo created by **jcomp** - Freepik.com

Text & Image slide photo created by **rawpixel.com** - Freepik.com

Text & Image slide photo created by **Freepik**