Exploring Text Classification for Enhancing Digital Game-Based Language Learning for Irish

Leona Mc Cahill[1†], Thomas Baltazar[1†], Sally Bruen[2], Liang Xu[1]
Monica Ward[1], Elaine Uí Dhonnchadha[2], Jennifer Foster[1]
[1]School of Computing, Dublin City University
[2]School of Linguistic, Speech and Communication Sciences, Trinity College, Dublin
{leona.mccahill2,thomas.baltazar2}@mail.dcu.ie,[†]Joint first authors
{liang.xu, monica.ward, jennifer.foster}@dcu.ie
{sbruen, uidhonne}@tcd.ie

Abstract

Digital game-based language learning (DGBLL) can help with the language learning process. DGBLL applications can make learning more enjoyable and engaging, but they are difficult to develop. A DBGLL app that relies on target language texts obviously needs to be able to use texts of the appropriate level for the individual learners. This implies that text classification tools should be available to DGBLL developers, who may not be familiar with the target language, in order to incorporate suitable texts into their games. While text difficulty classifiers exist for many of the most commonly spoken languages, this is not the case for under-resourced languages, such as Irish. In this paper, we explore approaches to the development of text classifiers for Irish. In the first approach to text analysis and grading, we apply linguistic analysis to assess text complexity. Features from this approach are then used in machine learning-based text classification, which explores the application of a number of machine learning algorithms to the problem. Although the development of these text classifiers is at an early stage, they show promise, particularly in a low-resourced scenario.

Keywords: text classification, under-resourced language, digital game-based language learning

1. Introduction

Language learning is a challenging process and is even more difficult when motivation levels are low. This is often the case with 'smaller' languages, including languages like Irish. Digital game-based language learning (DGBLL) tools can help in the language learning process, but they are difficult to develop. Often the developers are specialists in game development and not necessarily experts in linguistics or Computer Assisted Language Learning (CALL). For many well-resourced languages, the developers can avail of a variety of Natural Language Processing (NLP) tools to help them build DGBLL resources for these languages. For example, they can use text classifiers to determine suitable texts for students of different abilities (Crossley et al., 2023). However, for lesser-resourced languages, these tools may not exist and that makes it difficult to develop pedagogically suitable games for these languages.

This paper looks at the development of text analysis tools for Computer Assisted Language Learning (CALL), with a focus on less commonly taught languages (Irish in particular). The format of the paper is as follows. We provide a brief overview of NLP and CALL for Irish and of *Cipher* - a DBGLL application for Irish. We then describe our dataset and various Machine Learning approaches to the development of text difficulty classifiers for Irish. We report our results to date and conclude by pointing to future work in this area.

2. Background

2.1. NLP for CALL and Irish

NLP resources such as text analysers have the potential to contribute to Computer-Assisted Language Learning (CALL) but they remain largely under-used (Ward, 2019). This is because NLP focuses on language, linguistics and technology with limited consideration for pedagogy, whereas CALL researchers focus on pedagogy first and technology second. Therefore there is limited overlap between the two areas. As it is difficult to develop NLP resources, naturally there are fewer NLP resources for lower-resourced languages. This imposes an additional challenge to the use of NLP tools in CALL resources.

Although Irish is the first official language of Ireland, it is only spoken on a daily basis by less than 2% of the population (CSO, 2016). Therefore, there is a great need for additional sources of language input, such as games, for L2 learners. Irish is a compulsory subject in both primary and secondary schools in Ireland, but given that there is a very small number of learners on a worldwide basis, it is often not economically feasible for companies to develop Computer Assisted Language Learning

SIGUL2024 Workshop, pages 90–96

(CALL) resources for Irish.

2.2. Cipher Project: Context and Motivation

The Cipher project (Xu et al., 2022) explores the integration of a digital game into language learning, in this case targeting the Irish language. Cipher is a DGBLL game that leverages the engaging mechanics of gameplay to facilitate language learning, particularly in the context of endangered or low-resourced languages. Cipher emphasises pedagogical foundations while maintaining an enjoyable game design (see Fig. 1). It aims to address certain challenges in Irish language learning, such as orthographic complexity and learner motivation issues, by encouraging language learning through gameplay. The game's design incorporates sociocultural approaches, linguistic elements, and advanced technology to enhance comprehension and engagement. Feedback from learners and teachers has highlighted Cipher as a promising tool for language acquisition and cultural reconnection. An adaptive approach is used whereby texts may need to be of a higher or lower difficulty level depending on player characteristics and their performance in the game. It is important to ensure that the texts presented to the player are of a suitable level. This paper explores the development of text analysis tools for Irish which are necessary to enhance the educational outcomes of Cipher.

Figure 1: A screenshot of Cipher

2.3. Text Difficulty Classification

Text analysis and text grading has been a popular research area in linguistics as it can aid language learners to progress gradually by building their vocabulary and other language skills. Much of the research to date surrounding text analysis and text grading has been carried out on major languages such as English (Balyan et al., 2018; Ding et al., 2022; Pujianto et al., 2019) while languages such as Irish have not been researched to the same extent. Our goal is to apply the tools used for text grading and analysis in other languages to the Irish language. Previous research (Ó Meachair, 2019; Uí Dhonnchadha et al., 2022) shows that lexical and grammatical complexity play an important role in text grading for Irish. Therefore lexical, grammatical and frequency measures were calculated as input features to the ML models.

3. Dataset

3.1. Test Set

In order to build a text difficulty classifier for Irish, a suitable dataset must be built, since none currently exist for the Irish language. To create our dataset, we need to collect as much labelled Irish text data as is publicly available across the internet. We decided to mainly focus on two websites: `ccea.org.uk` which is an Irish language resource for schools in the UK and `scoilnet.ie` which is a primary and post-primary school website which contains Irish resources for different class groups. Texts from each of these websites were extracted along with their respective labels that can be used to predict the class (grade) range for a sample of Irish text across primary and secondary school level. We decided on 5 levels, with 1 representing 1st-2nd class (ages 6-8), 2 representing 3rd-4th class (ages 8-10), 3 representing 5th-6th class (ages 10-12), 4 representing lower secondary/middle school level (ages 12-15) and 5 representing upper secondary/high school level (ages 15-18). This test set consists of 190 labelled non-translated Irish text samples from the two websites `ccea.org.uk` and `scoilnet.ie`. It also contains some manually labelled Irish stories used in the Cipher game mentioned above.

3.2. Training Set

Since there was not enough labelled Irish data across these websites to train an effective ML model we explored other options to get more training data, in particular machine translation of existing labelled text datasets for the English language. One such publicly available dataset is Clear Corpus (Crossley et al., 2023)[1], which contains thousands of English text excerpts, with various difficulty metrics calculated on each. There are texts in different genres such as fiction, history, science and poetry, with a combination of different difficulty scores such as the Automated Readability Index and Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975), as well as the Crowdsourced Algorithm of Reading Comprehension (CAREC) (Crossley et al., 2019) and the Coh-Metrix L2 Readability Index (Crossley et al., 2008). The Clear Cor-

[1] https://github.com/scrosseye/CLEAR-Corpus

Dataset Split	Total Samples	Source
Train	2610	Clear Corpus (translated), chatgpt
Validation	653	Clear Corpus (translated), chatgpt
Test	190	ccea.org.uk, scoilnet.ie, cipher

Table 1: Dataset Statistics

pus also contains a unique difficulty metric called BT_easiness (Bradley and Terry, 1952) which was calculated using manual rankings by teachers, who were given two texts and asked to rank which one was more difficult.

The first step in making this dataset useful for our project was to translate each of the 3195 excerpts to Irish, using the Google Translate library in Python. We did this with the assumption that the translations were mostly accurate and that a more difficult English text translated to Irish would be more complicated than a simpler English text translated to Irish, i.e. the difficulty labels would be preserved.

Once the text was translated, we needed to use the different difficulty labels to create an overall level that corresponds to the levels 1-5 mentioned above for Irish L2 school learners, which probably will not coincide with the L1 English grading. We first looked at the given lexile level assigned to the respective English texts to see how many texts there were at each different grade level. We realised most of the texts were at higher grade levels 9th grade + (level 5) and there were not many texts at the lower grades (level 1). We then mapped the BT_easiness, L2 Readability Index and lexile level scores to an appropriate level 1-5. An average of these three levels was calculated to get an overall level which was rounded to the nearest whole number. To validate how accurate the levels were for Irish we calculated some automatic difficulty measures used in the Clear Corpus on the Irish translated text. We calculated FKGL and Automated Readability Index on the Irish text and converted these grade scores to our levels 1-5. We then compared this to our BT_easiness, lexile level and L2 Readability Index average level, and found a good overlap. We then incorporated these scores into the calculation of the final level label. One was added to each label as these scores assumed Irish as a first language whereas for most students across the country that is not the case. When consulting Irish primary school teachers they recommended this increase and said that the easiest text in the dataset would probably be too challenging for most 1st and 2nd class students, which resulted in data labelled 2- 6 to be used for training.

To get Irish data for 1st-2nd class students for use in training our model we had to find another text source. After finding some basic 1st- 2nd class level sentences on the web we used these to prompt chat-gpt[2] to generate more text excerpts. We looked over each of these generations, making changes and deletions where necessary. Ultimately we were able to add 180 level 1 (1st-2nd class) excerpts to our training set. The training set was then split to create a validation set for the models. This resulted in 2610 entries in the training set, and 653 rows in the validation set – see Table 1.

4. Methodology

4.1. Baseline Features

This method involves calculating linguistic measures specifically for Irish on pre-graded data and using these measures as features to predict the difficulty levels. To investigate the most useful linguistic measures for Irish texts, pre-graded texts for use in Irish primary schools were used. Stories from Séideán Sí (SS) and Taisce Tuisceana (TT) were sourced on www.cogg.ie. Various lexical and grammatical measures were calculated for this data set (Vajjala and Meurers, 2012). For this data, the lexical measures TTR (type token ratio), WTR (word type ratio) and CTTR (corrected type token ratio) as well as grammatical measure WDSEN (average number of words per sentence) appeared to be best at distinguishing between each age group showing an increase between 1st to 6th class stories, as shown in Figs. 2 and 3. These

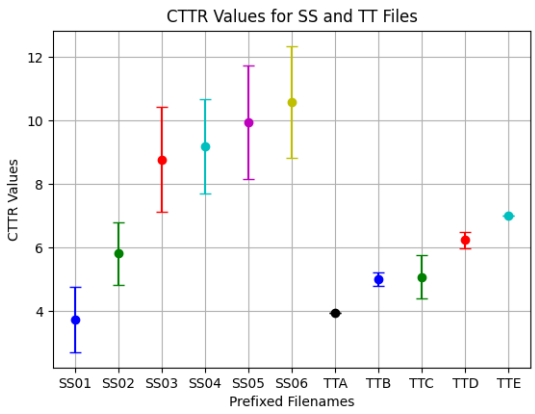

Figure 2: CTTR values for Séideán Sí (SS) and Taisce Tuisceana (TT) texts

[2]https://chatgpt.com/, accessed 19th January 2024

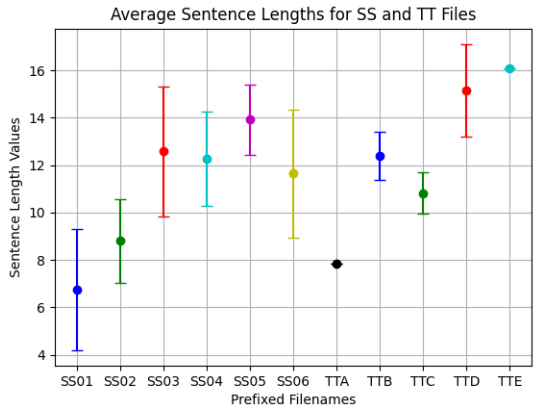

Figure 3: Avg. Sentence Length values for Séideán Sí (SS) and Taisce Tuisceana (TT) texts

Model	Val	Test
LR Baseline Features	62	41
LR TFIDF w SMOTE	56	43
LR TFIDF w/o SMOTE	52	42
RR TFIDF w SMOTE	56	41
RR TFIDF w/o SMOTE	55	41
mBERT	77	40
gaBERT	**80**	31
bi-LSTM	54	**50**
CNN	51	47

Table 2: Classification Accuracy on the Validation and Test Sets. LR: Logistic Regression. RR: Ridge Regression.

4 metrics were then calculated on our training and test data and used as the baseline features to train our model. A basic autoML experiment was run on the training and validation data using Pycaret and it was found Logistic Regression performed the best.

4.2. Classification with Traditional ML

Features The features used in the traditional ML experiments are Tf-Idf-weighted word counts. Tf-Idf features take into account the frequency of a word in a document in proportion to the amount of documents overall that the word occurs in. To prepare the texts for Tf-Idf vectorisation, stop words were removed (using a custom made list for Irish) and words were lowercased.

Algorithms Before deciding on which multiclass classification algorithms to use, a basic autoML experiment was run on the training and validation data using Pycaret. In order to determine if accuracy of the classification algorithms would be higher when trained on a set of balanced classes, we experimented with oversampling using Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002).

The top four performing models were chosen for manual experiments. The four models were trained on two versions of the training data: the original version and the SMOTE oversampled version. The four classification models used for the experiment were the ridge regression, logistic regression, extreme gradient boost (XGBoost) and random forest classifiers.

4.3. Neural Network Classification

Features The default Tokenizer class in tensorflow was used to vectorize the text. The input text was split into individual words or tokens, with unique words were mapped to integer indices.

Algorithms We experimented with deep learning models in the form of neural networks in an attempt to capture more contextual information and non-linear relationships in our data. Recurrent Neural Networks including uni- and bi-directional LSTMs were tried, as well as Convolutional Neural Networks (Hochreiter and Schmidhuber, 1997; Kim, 2014). We experimented with the number of layers, embedding dimension size and learning rate to find the parameters that worked best for our data.

4.4. Pretrained Language Models

As well as traditional ML classification, experiments were run to investigate the performance of pre-trained neural language models on the text difficulty classification task. We fine-tuned language models that have been pretrained on multilingual data and/or Irish data. Two language models were used – multilingual BERT (Devlin et al., 2019) and monolingual gaBERT (Barry et al., 2022). Multilingual BERT was pre- trained on Wikipedia text with 104 different languages, and the gaBERT model was pre-trained solely on Irish text, including Irish language Wikipedia text, the Irish side of English-Irish parallel corpora and the National Corpus of Ireland (Kilgarriff et al., 2006). When tokenising the text for the gaBERT model, the maximum padding length was set to match the maximum length of the multilingual BERT model. The performances of the models were measured based on training/validation loss and validation accuracy. Both models were trained for 3 epochs.

5. Results

Table 2 summarises the different classification algorithms and language models used, along with their accuracy scores against the validation set and

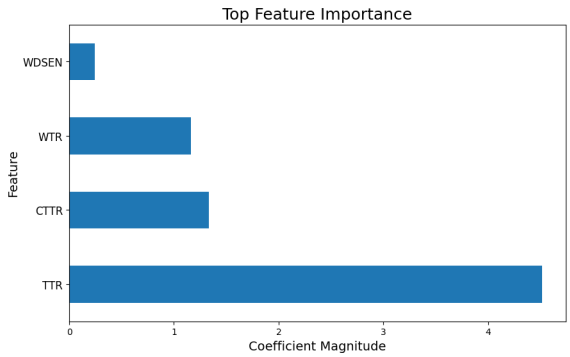

Figure 4: Baseline features: relative importance

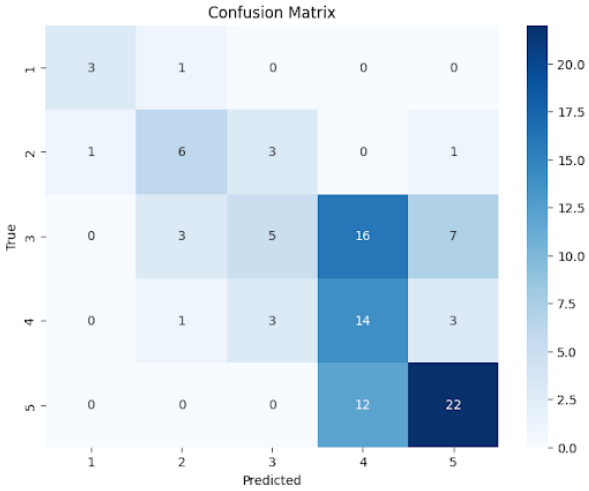

Figure 5: bi-LSTM Confusion Matrix for Unseen Data (Scoilnet only)

unseen test set.[3]

For all models, we observe that there is a substantial difference in accuracy between the validation and test sets. This trend can be explained by the fact that the validation set texts have been translated from English or, in the case of the simpler text, generated by a large language model, whereas the test set texts are Irish-language text used to teach Irish. The best performing approach on the test data was the bi-LSTM neural network, followed by CNN. The best models to choose when the training/test data align are the fine-tuned language models (gaBERT and multilingual BERT) since these are the top performing models, by a large margin, on the validation data. However, this performance did not translate to the unseen data, highlighting the substantial differences between the train/validation and the test data.

The test data comes from two sources: `ccea.org.uk` and `scoilnet.ie`. The Logistic Regression model with baseline features performed better on documents from CCEA, whereas this was not the case for the bi- LSTM classification. Feature importance for the Logistic Regression model with baseline features was found by retrieving the absolute coefficient value for each feature. Fig. 5 shows that the most important baseline feature in determining the difficulty of texts in Irish was the type-token ratio. The difficulty classification was influenced the most by the lexical diversity of the sentences.

Fig. 5 depicts the confusion matrix of the bi-LSTM network on the Scoilnet subsection of the unseen data. The model performed the best in classifying texts of difficulty level 5. The network confused texts of level 3 with those of level 4, as well as level 5 text exhibiting similar traits to level 4 text.

6. Conclusion

There is a need for NLP tools such as text classifiers for low-resource languages, which can help DGBLL developers select suitable texts for language learners. In this paper, we have outlined a series of machine learning experiments on the task of text difficulty classification for Irish. Predictive features were developed based on text analysis of pre-graded Irish resources, and a variety of classification algorithms were tried, including classical and neural approaches as well as neural language model fine-tuning.

The current results, although promising, are preliminary and further tests will be carried out on more unseen data. We aim to increase the amount of Irish texts that can be used in model training and to improve data quality by seeking the help of primary school teachers to manually assign a difficulty level to the texts. Future work also involves improving the classification models so that they may be at an adequate enough standard to be implemented in the Cipher game. The aim would be to use the models to help the game ensure Irish texts are of a suitable difficulty level to assign to different age groups.

7. Acknowledgements

We thank the reviewers for their helpful feedback. This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

8. Bibliographical References

[3]Note that only the top-performing models from the ML and neural network groups are included.

Renu Balyan, Kathryn McCarthy, and Danielle McNamara. 2018. Comparing machine learning classification approaches for predicting expository text difficulty. In *Proceedings of the Thirty-First International Florida Artificial Intelligence Research Society Conference, pp.*, pages 421–426.

James Barry, Joachim Wagner, Lauren Cassidy, Alan Cowap, Teresa Lynn, Abigail Walsh, Mícheál J. Ó Meachair, and Jennifer Foster. 2022. gaBERT — an Irish language model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4774–4788, Marseille, France. European Language Resources Association.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

S. Crossley, A. Heintz, J.S. Choi, J. Batchelor, K. Mehrnoush, and A. Malatinszky. 2023. A large-scaled corpus for assessing text readability. *Behav Res*, 55:491–507.

Scott A. Crossley, Jerry Greenfield, and Danielle S. McNamara. 2008. Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42(3):475–493.

Scott A Crossley, Kristopher Kyle, and Mihai Dascalu. 2019. The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior research methods*, 51:14–27.

CSO. 2016. Census of population 2016. Accessed on 2023-02-20.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.

Han Ding, Qiyu Zhong, Shaohong Zhang, and Liu Yang. 2022. Text difficulty classification by combining machine learning and language features. In *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery*, pages 1055–1063, Cham. Springer International Publishing.

Arthur C. Grasser and Danielle S. McNamara. 2011. Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, 3:371–398.

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Adam Kilgarriff, Michael Rundell, and Elaine Uí Dhonnchadha. 2006. Efficient corpus development for lexicography: building the New Corpus for Ireland. *Language Resources and Evaluation*, 40:127–152.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

D. Malvern, B. Richards, N. Chipere, and P. Durán. 2004. *Lexical Diversity and Language Development: Quantification and Assessment*. Springer.

M. J. Ó Meachair. 2019. *The Creation and Complexity Analysis of a Corpus of Educational Materials in Irish (EduGA)*. Ph.D. thesis, Trinity College, Dublin.

Utomo Pujianto, Muhammad Fahmi Hidayat, and Harits Ar Rosyid. 2019. Text difficulty classification based on lexile levels using k-means clustering and multinomial naive bayes. In *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pages 163–170.

Elaine Uí Dhonnchadha, Monica Ward, and Liang Xu. 2022. Cipher – faoi gheasa: A game-with-a-purpose for Irish. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 77–84, Marseille, France. European Language Resources Association.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop*

on Building Educational Applications Using NLP, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

Monica Ward. 2019. *Joining the blocks together – an NLP pipeline for CALL development*, pages 397–401.

Liang Xu, Elaine Uí Dhonnchadha, and Monica Ward. 2022. Faoi gheasa an adaptive game for Irish language learning. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 133–138, Dublin, Ireland. Association for Computational Linguistics.