Language Identification with Character N-grams

November 22, 2023

1 Introduction

This report presents the results of a language identification task using character-based N-grams for a Czech-Polish language dataset. We discuss the methods, classification results, and linguistic insights derived from this task.

2 Method

Calculate the frequency for each ngram in the training dataset and stores it in separate dictionaries for each language
For each sentence in the test dataset, the algorithm breaks down the sentence into n-grams of size n and checks each n-gram against the n-grams dictionaries of Polish and Czech
Compares their frequencies and increases the counter for the language with the higher frequency. After analyzing all n-grams in the sentence, the algorithm gives the language with higher counter value as the predicted language.

2.1 Classification Metrics

We calculated the precision and recall for each N-gram class, as shown in Table 4.

Table 1: Precision and Recall for N-gram Classes

N-gram Class	**Precision**	**Recall**
1gram	0.7700569739789548	0.6415094339622641
2gram	0.7998873629680612	0.7091125703044912
3gram	0.9474497866365013	0.9427866899398776
4gram	0.987616706384878	0.9874213836477987

3 Misclassified Sentences

Some misclassifed sentences are due to the following factors: 1) Unbalanced data - The number of sentences in the training data set is more for Polish (93437) than Czech (50932) so expectedly, the frequency of ngrams for Polish will be more than Czech

Table 2: Precision and Recall for N-gram Classes after normalisation

N-gram Class	Precision	Recall
1-gram	0.7959109657652166	0.5372509904967723
2-gram	0.9794780573595491	0.9793865846563046
3-gram	0.9925201476894107	0.9925193250768847
4-gram	0.996046220186964	0.9960380129110908

2) Few names and Proper nouns appear in the training data of one language more than the other this creates Eg Richard Roberts jest autorem licznych książek
2) Borrowed words like internet, Fantastické also cause misclassification of sentences

4 Note on Zero Count N-grams

In this test set, all the n-grams were present in at least one of the languages. So assuming if an n-gram is present in only one of the languages, that sentence belongs to that language only, on putting this constraint on the model, the results are significantly better for unigrams but worse for trigrams and 4-grams. This is due to the number of permutations for unigrams being way lesser than trigrams and 4-grams therefore, their distribution is more accurate. The accuracy is expected to increase if the dataset is large as seen in Spanish and Portuguese.

Table 3: Precision and Recall for N-gram Classes after handling zerp

N-gram Class	Precision	Recall
1gram	0.9354931908409589	0.92170226913806
2gram	0.9888574712390337	0.9887512814119082
3gram	0.9801764721324643	0.9792757598426287
4gram	0.9362192278119122	0.922699692461142

However, in any test set, one can encounter cases where certain N-grams had zero counts for both languages, this can be problematic for classification, as the model may not reliably predict those N-grams.

5 Successfully Classified Sentences

In cases where the system successfully classified sentences, we observed that character N-grams captured various linguistic phenomena specific to Czech and Polish. This included orthographic differences, characters, word endings, vowel harmony, and common prefixes and suffixes.

6 Bonus Spanish and Portuguese

Results and Observations for this data set is also similar. Precision and Recall for Language identification using the method explained

Table 4: Precision and Recall for N-gram Classes

N-gram Class	Precision	Recall
1-gram	0.7005732637213983	0.5213187128029484
2-gram	0.876277375334656	0.8761118418482295
3-gram	0.9511803643601784	0.9509943089325016
4-gram	0.9761608256581925	0.9761027257218019

7 Conclusion

The report demonstrates the effectiveness of character N-grams in capturing language-specific features for the Czech-Polish language pair and Spanish-Portuguese language pair. However, challenges arise with mixed sentences and borrowed words, Proper Nouns, and understanding the system's limitations is crucial for reliable classification.