Interchangeability of ngrams models between heterogeneous dataset.

Cuper, Mirjam
mirjam.cuper@kb.nl
KB, national library of the Netherlands, Netherlands, The

Introduction

A lot of heritage material is digitally available with more being added continuously. Regrettably, OCR software used for digitizing texts is not flawless. This leads to a variety of quality in digitized texts. In order to solve these problems, OCR errors must be detected and resolved.

A method that appears to be an accurate way for the detection and correction of errors in (digitized) texts is the use of ngram models. For example, ngrams models are used to detect spelling errors (El Atawy / Abd ElGhany 2018; Wu et al. 2013) and are implemented in the quality pipeline of both the National Library of the Netherlands and the National Library of Luxembourg (Cuper 2022; Schneider / Maurer 2022). Furthermore, various studies use ngram models for the post-processing of OCRed texts (Nguyen et al. 2021; Chiron et al. 2017).

However, there is a huge variety of types of heritage texts that are digitally available. Not only are there different types of publications, such as fiction and non-fiction books, periodicals, and newspapers, but time periods also need to be taken into consideration. Due to this large variation in material, and because of the effort to create manually corrected versions of texts, representative ngrams models are only available for a small number of corpora.

It is not evident that ngrams models built upon one corpus can be interchanged to measure the quality of another corpus with different characteristics. Previous research has shown that ngrams can be used to measure similarity between texts (Islam et al. 2012). Furthermore, ngrams have been successfully applied for author-profiling (Basile 2017). These results imply that there are significant differences between ngram models of various corpora. Following this line of thought, each type of corpus may very well need their own ngram model in order to use ngram models for error detection and correction. This would mean that using ngram models is only feasible when there is an ngram model of a representative corpus available.

This leads to our research question: *How, and to what extent, is an ngram model built upon corpus A interchangeable to measure the quality of corpus B.*

Methodology

We propose to determine how interchangeable ngram models are when applied beyond the scope of their original corpus. We focus on two separate corpora combinations (both will be presented on the poster):

- The interchangeability of ngram models from various reading levels.
- The interchangeability of ngram models from various time periods.

To compare corpora of various reading levels, we use a combination of standard English Wikipedia articles (7.649.864 words) and articles from simple English Wikipedia (7.648.884 words). For the comparison on time periods, we use a set of Dutch books from 1584 to 1883 (1.893.137 words) and a set of more modern books from 1947 to 1999 (1.993.072 words). For both combinations, we will create both word- and character ngram models on a train set, with ngrams ranging from bigrams to fourgrams. The ngrams and their corresponding frequencies are created using the Python package NLTK, after which this information was used to calcuate the log probability. To calculate the similarity between the various models, a wilcoxon signed-rank test was used. Only ngrams with a minimal frequency of 10 were used in this analysis.

To get a broader understanding of the meaning of the results from the similarity test, we perform some additional tests: we perform a similarity test between ngram models created from two samples from the same corpera, we analyse the intersection of unique ngrams between models, and we will evaluate the models using intrinsic evaluation, in which the perplexity of our models on the test sets will be calculated

Results

Our preliminary results based on the various word ngrams models can be found in table 1 and 2. Except for the log probability of the bigram model, we found significant differences between the ngrams models of the various corpera. Our poster will be supplemented with the results from our analyses on the character level and the results from our intrinsic evaluation. To support our findings the characteristics of the used datasets will be shown and the implications for measuring the quality of digitized texts will be discussed.

Table 1: Wilcoxon signed-rank test results, significant p-values in bold.

	simpel versus standard Wikipedia	
	p-value Frequency	p-value Log probability
Bigram	$1.25x10^{-70}$	0.34
Trigram	0.0	0.0
Fourgram	0.0	0.0

Table 2: Wilcoxon signed-rank test results, significant p-values in bold.

	Books from 1584-1883 versus books from 1947-1999	
	p-value Frequency	p-value Log probability
Bigram	$4.39x10^{-10}$	0.86
Trigram	$2.94x10^{-28}$	$6.87x10^{-43}$
Fourgram	$1.11x10^{-6}$	$5.05x10^{-11}$

Bibliography

Aminul, Islam / Evangelos, Milios / Vlador Kešelj (2012). “Text similarity using google tri-grams.” in: *Proceedings of the 25th Canadian conference on Advances in Artificial Intelligence (Canadian AI'12)*. Springer-Verlag, Berlin, Heidelberg, 312–317. DOI: *https://doi.org/10.1007/978-3-642-30353-1_29 ’*

Basile, Angelo/ Dwyer, Gareth / Medvedeva, Maria / Rawee, Josine / Haagsma, Hessel / Nissim, Malvina (2017). “N-GrAM: New Groningen Author-profiling Model.” in: *Working Notes of*

{CLEF} 2017 - Conference and Labs of the Evaluation Forum, 1866

Chiron, Guillaume / Douce, Antoine / Coustaty, Mickaël / Moreux, Jean-Philippe (2017). "ICDAR2017 Competition on Post-OCR Text Correction *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1423-1428, DOI: 10.1109/ICDAR.2017.232.

Cuper, Mirjam (2022). "Examining a Multi Layered Approach for Classification of OCR Quality without Ground Truth" in: DH-Benelux journal Volume 4, The Humanities in a Digital World, DOI: *https://doi.org/10.17613/03ds-9973*

El Atawy, S. / Abd ElGhany, A. (2018). "Automatic spelling correction based on n-gram model." In: *International Journal of Computer Applications*, 182:5–9, DOI: 10.5120/ijca2018917724.

Nguyen, Thi Tuyet Hai / Jatowt, Adam / Coustaty, Mickael / Doucet, Antoine (2021). "Survey of Post-OCR Processing Approaches" in: Association for Computing Machinery. 54. DOI: https://doi.org/10.1145/3453476

Schneider, Pit / Maurer, Yves (2022). "Rerunning OCR: A Machine Learning Approach to Quality Assessment and Enhancement Prediction" in: *Journal of Data Mining & Digital Humanities*. DOI: 10.46298/jdmdh.8561

Wu, Jian-Cheng / Chang, Jim / Chang, Jason J. S. (2013). "Correcting serial grammatical errors based on n-grams and syntax." in: *Int. J. Comput. Linguistics Chin. Lang. Process.*, 18.