Revisiting CCNet for quality measurements in Galician

John E. Ortega*[0000−0002−2328−3205], Iria de-Dios-Flores[0000−0002−5941−1707], José Ramom Pichel[0000−0001−5172−6803], and Pablo Gamallo[0000−0002−5819−2469]

Centro de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
{john.ortega,iria.dedios,jramon.pichel,pablo.gamallo}@usc.gal

Abstract. In this article, we report the construction of a web-based Galician corpus and its language model, both made publicly available, by making use of CCNet tools and data. An in-depth analysis of the corpus is made so as to provide insights on how to achieve optimum quality through the use of heuristics to lower the perplexity.

Keywords: Galician · Web Crawl · Quality Estimation · Perplexity

1 Introduction

In recent years, there has been a higher need for quality data to create high performing machine learning models that typically use neural networks. This has especially been the case for languages that have low to medium resources where approaches require an immense amount of digital text. The varieties of such languages that we are concerned with are the Galician-Portuguese branch, which belongs to the Western Ibero-Romance group, spoken in the northwest of the Iberian Peninsula. This article highlights the efforts of a new proposal to increase the use of Galician within a novel project called *Nós*[1], starting off with the first gathering of the CCNet corpus. We report the first of a series of tasks that will attempt to provide higher quality texts in Galician for training both supervised and unsupervised statistical models that will, in turn, be used for various natural language processing tasks. Our work introduces the CCNet corpus for Galician and makes it publicly available for anyone to download in an easy way. Additionally, we provide evidence from a systematic human-based analysis carried out by Galician expert reviewers that the original CCNet corpus contains several caveats. We document those caveats and describe the analysis performed on the CCNet quality. In this article, we report three main contributions: (1) A Galician corpus made available publicly, (2) a Galician language model also made publicly available, and (3) an in-depth analysis of the corpus

* Author to whom correspondence should be addressed.

[1] https://nos.gal

that provides insight on how to achieve optimum quality through the use of heuristics to lower the perplexity for use in natural language processing (NLP) tasks in future work.

2 Background and related work

In this article, we present our findings on reproducing the introduction of the common crawl corpus by Facebook, known as the CCNet corpus [11]. The CCNet corpus attempts to comprise monolingual data for 174 languages using a more common corpus known as the Common Crawl[2] corpus based on a "snapshot" of the web at a given time – the work on CCNet is based on the Common-Crawl dataset from February 2019 which is composed of 1.5 billion documents in 174 languages. Documents and languages in the CCNet work as presented on GitHub[3] and in the original paper are limited to a small subset of languages. Statistics are provided by CCNet [1] for 25 of the 174 languages and downloadable language models are only available for 48 languages (see Makefile from GitHub[3]). Unfortunately, language models for Galician and a downloadable corpus is not readily available, despite other efforts found online[4] that publish downloadable corpora for several of the missing languages yet do not claim authorship.

We present the CCNet corpus for Galician after processing and also freely provide the language model[5]. This forms part of an initiative to provide more Galician text and work available for wide use in an effort to increase visibility and add authorship to future projects based on the CCNet corpus for Galician. To this effect, we believe that the analysis presented here is unprecedented and will help establish some of the caveats of using the CCNet corpus for localized low-to-medium resource languages that are not included in the 25 to 48 language models currently available for use.

3 Methodology

This work is the first step of several in a multi-part project on Galician[6]. The project will help increase quality for several NLP tasks including machine translation, information extraction, text generation, and more. Since the CCNet corpus is most probably used for unsupervised machine learning (see [11] for an example of an unsupervised implementation), our plan is to create the highest quality corpus possible while at the same time gathering as many sentences, or segments, as possible. This was done by compiling a large Galician corpus

[2] https://commoncrawl.org/about
[3] https://github.com/facebookresearch/cc_net
[4] http://data.statmt.org/cc-100
[5] https://github.com/proxecto-nos/propor2022
[6] https://www.lingua.gal/recursos/todos/_/promovelo/contido_607/nos-intelixencia-artificial-servizo-lingua-galega

of high-quality segments that have been improved using several heuristics. Our method comprised the following tasks: (1) use the CCNet corpus project tools from GitHub to download the de-duplicated Galician text as described in the CCNet article[11]; (2) compile the CCNet corpus and make the de-duplicated text available for public use; (3) create a Galician language model based on KenLM[4] after tokenizing with SentencePiece[6] as described in the CCNet article and make it available for public use; (4) select 1000 random sentences for a a systematic qualitative analysis carried out by two Galician expert reviewers; (5) define several heuristics that can be used to better the already de-duplicated CCNet corpus in turn lowering the perplexity and apply them on 111 sample documents to verify their usefulness by calculating the perplexity before and after applying the heuristic.

Since the de-duplicated corpus and language model are not readily available on-line, the bulk of the work to be performed in order to create the high-quality corpus is related to the downloading of the corpus and creation of its model. On a 2.20 GHz Intel(R) Xeon(R) Silver 4214 CPU, approximately 2,000 documents can be processed in one day using the "debug" execution mode (single-threaded and single-process execution) from CCNet since parallel processing uses Slurm[7] which does not work well in our experiments[8]. Since there are a total of 64,000 crawls to be processed[9], the download and de-duplication process takes nearly a month. After that, the SentencePiece tokenization and KenLM model creation require less time (nearly 5 days on the same architecture) since the CCNet tools include code for training the model.

1000 random segments are chosen and given to two native Galician speakers. Both speakers are trained in Galician linguistics and have some background on NLP. They are asked to mark the original positions of text and provide what text should be modified and how. Patterns are then searched from both annotations which resulted in the heuristics described in the next section.

The final goal for this work is to eventually attain a high-quality Galician corpus. We base our definition of quality on *perplexity* as described in previous work on isolated European languages [3]. In the current work, we use it to measure the distance between noisy and cleaned texts of the Galician corpus. Our word tri-grams perplexity model was trained on Wikipedia data sets similar to previous work [3] for Galician. Tools used to create the model can be found on GitHub.[10]

4 Analysis and results

We report the error analysis, the heuristics applied to improve the quality of the text and the perplexity scores comparing the two data-sets. After de-duplication,

[7] `https://slurm.schedmd.com/`

[8] Similar issues have been reported for other languages at `https://github.com/facebookresearch/cc_net/issues`

[9] Each crawl consists of several documents totalling around ≈150 Megabytes in size.

[10] `https://github.com/proxecto-nos/propor2022`

the corpus contains 4,100,006 tokens and 6,496,871 lines.[11] As an aside, the original size reported by CCNet is of 440 MB and around 400k documents, our downloaded corpus is the same size (440 MB), but the number of documents in our experiment is 64k compared to 400k in the CCNet article, this is due to the CCNet system automatic break up of the documents to parallelize processing. The final results left around 11% of the original tokens and less than half of the number of characters – a drastic reduction of the original crawled corpus.

4.1 Error Analysis

Several issues were identified by the two Galician expert reviewers in our experiments. One of the recurring issues, as stated from the original CCNet work [11], was the introduction of other languages, namely English and Spanish. Thus, much like other research on low-resource languages [7, 9, 10], we found that the higher-resource neighbor, Spanish, was often present within the Galician texts.

The main language error types found in the CCNet de-duplicated corpus were the following: (1) English and Spanish excerpts ranging to as many as 80 words long; (2) book publications along with their titles, authors, and other bibliographical information; (3) bibliographical entries containing mundane punctuation or long alphabetically-sorted lists separated by delimiters; (4) misspelled or non-standarized Galician words; (5) boilerplates and typical errors from web crawls such as the inclusion of HTML tags, Javascript, the combining of two or more words into one, and other non-human readable text.

4.2 Heuristics

In order to mitigate the issues found during error analysis, we propose four heuristics for lowering the perplexity of the Galician corpus. These heuristics cover the main issues discovered during error analysis but further discovery will be considered in future work.

Heuristic 1. Remove non-Galician words since there are several words in English, Spanish, and other languages. **Heuristic 2**. Words in lists in any language can be removed or broken down into separate words (or tokens). **Heuristic 3**. Splitting of combined words can be performed by using character-level splits along with SentencePiece [6] tokenization and byte-pair encoding [5]. **Heuristic 4**. Replace those words that have been misspelled by making use of a Levenshtein-based [8] character distance between in-vocabulary words found in a Galician word dictionary. An initial threshold such as 90% could be used to measure the distance between the misspelled word and the potentially correct words.

In order to show that the implementation of heuristics could lower complexity of the CCNet Galician corpus, we test Heuristic 1 by removing non-Galician text and measuring the quality of the text before and after the removal by using

[11] In the original CCNet article, the sentence splitting uses Moses but the version is not clear, we report the number of lines here instead.

perplexity. Non-Galician text was removed by making use of QueLingua language detector [2], software originally tuned for Galician. Our results show that the average perplexity of 111 documents before removing the text is 5028.054 and after removing the text is lower, 4988.369. While initially, this is not a huge gain, we note that this was done for 111 documents only. Future work will apply all four heuristics over the entire corpus.

5 Conclusion and Future Work

We believe that the inclusion of all four heuristics introduced over the entire corpus will result in significantly lower perplexity. We have shown that with a small sample of documents the most problematic issue can be resolved to increase the quality. We have downloaded the CCNet Galician corpus and made it publicly available in an easy-to-use format along with its corresponding language model. Future work will include the implementation and evaluation of the other three heuristics proposed in this work along with other heuristics. We also plan on evaluating the model and using both the corpus and model in various NLP tasks for Galician and Portuguese.

6 Acknowledgements

This research was funded by the project "Nós: Galician in the society and economy of artificial intelligence", agreement between Xunta de Galicia and University of Santiago de Compostela, and grant ED431G2019/04 by the Galician Ministry of Education, University and Professional Training, and the European Regional Development Fund (ERDF/FEDER program).

References

1. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.747, https://aclanthology.org/2020.acl-main.747
2. Gamallo, P., Alegria, I., Pichel, J.R., Agirrezabal, M.: Comparing two basic methods for discriminating between similar languages and varieties. In: COLING Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3) (2016)
3. Gamallo, P., Pichel, J.R., Alegria, I.: Measuring language distance of isolated european languages. Information **11**(4) (2020). https://doi.org/10.3390/info11040181, https://www.mdpi.com/2078-2489/11/4/181
4. Heafield, K.: Kenlm: Faster and smaller language model queries. In: Proceedings of the sixth workshop on statistical machine translation. pp. 187–197 (2011)
5. Kida, T., Fukamachi, S., Takeda, M., Shinohara, A., Shinohara, T., Arikawa, S.: Byte pair encoding: a text compression scheme that accelerates pattern matching (1999)

6. Kudo, T., Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226 (2018)
7. Lankford, S., Afli, H., Way, A.: Machine translation in the covid domain: an english-irish case study for loresmt 2021. In: Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021). pp. 144–150 (2021)
8. Levenshtein, V.: Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady. **10**(8), 707–710 (1966)
9. Ortega, J.E., Castro-Mamani, R.A., Samame, J.R.M.: Love thy neighbor: Combining two neighboring low-resource languages for translation. In: Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021). pp. 44–51 (2021)
10. Ortega, J.E., Mamani, R.C., Cho, K.: Neural machine translation with a polysynthetic low resource language. Machine Translation **34**(4), 325–346 (2020)
11. Wenzek, G., Lachaux, M.A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., Grave, E.: CCNet: Extracting high quality monolingual datasets from web crawl data. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 4003–4012. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.494