DSAN5400 - Final Project Proposal

Researching OSCAR Dataset Creation for Fact Verification Purposes

Group 7: Alex Choi, Shawn Xu, JaeHo Bahng, Michael Varnerin

Problem:

With the widespread use of the internet and social media, the amount of raw text information and data spreading online is unfathomably large and will continue to grow at a faster rate. This increasing amount of information generated on the internet also means that there is an increasing amount of misinformation being generated. Therefore, the process of extracting raw text data from the internet has the risk of containing false information, which would further complicate Natural Language Processing tasks. Furthermore, this challenge expands when considering the multilingual nature throughout the internet since misinformation from one language can easily be translated into multiple languages.

Our team aims to research this issue of detecting misinformation by analyzing multilingual corpora from the OSCAR dataset. Misinformation can shift public opinions, impact company decisions, and put public health at risk. For instance, there was a rise in misinformation involving vaccines and treatment methods during the COVID-19 pandemic, which greatly impacted public health decisions. By analyzing the OSCAR dataset, our team hopes to learn how to create multilingual datasets that will aid in developing models that can detect misinformation patterns across languages and different cultural contexts. Understanding, analyzing, and potentially improving on the methods from the OSCAR dataset can advance innovations in verifying facts from large amounts of collected, and processed data.

Current Research:

Two of the most prominent examples of high-quality corpora are The Pile and The RefinedWeb, which will both be referenced throughout this project. Throughout the creation of both The Pile and RefinedWeb, different methods to scrape, filter, and refine millions of rows of text data were used. For example, within the paper about The Pile, jusText was utilized to yield higher-quality data from the Common Crawl data set [1]. This tool, which was used by Endredy and Novak, splits HTML content into paragraphs using block-level tags and then classifies the content into buckets indicating the content's quality. [2] Another method used in prior research includes deduplication, which was used in the creation of RefinedWeb. Deduplication works to remove redundant text from a corpus, allowing for the creation of different pipelines to filter out other unwanted data [3]. However, while these datasets do filter for redundant data, they do not necessarily filter for false information nor do they contain annotations that indicate the language characteristics of a given line of data to prospective researchers, meaning that future researchers need to take the arduous step of cleaning and annotating this data for their projects, especially if they plan to train a multilingual model. To fill this gap, a dataset named OSCAR, or the Open

Super-large Crawled Aggregated Corpus, was created [4]. This project will analyze how the OSCAR dataset was created by evaluating the pipelines used to obtain, clean, and annotate the text data used to make it. Additionally, this project aims to include replication demos showing these pipelines in action to understand better how they can be used in future text-based projects.

The first major pipeline used to create the OSCAR dataset is the "goclassy" pipeline. This pipeline is an asynchronous pipeline that is based on the older fastText linear classifier [5]. Goclassy employs scaled parallelization to asynchronously open and process data based on the available computing resources on the current machine - as compared to the more serialized fastText pipeline. This allows for more efficient data processing as it is no longer necessary to finish working on one text file before moving on to the next, in addition to providing more machines with the ability to process this data given goclassy's scalability.

The second major pipeline, and the pipeline that builds off of goclassy, is the "Ungoliant" pipeline. The Ungoliant pipeline was first proposed in 2021 with the goal of further optimizing the data processing pipeline that goclassy introduced [6]. It achieves this primarily through the use of multithreading and avoiding the use of intermediate files - which goclassy heavily relied on. Additionally, Ungoliant implements many of these more aggressive time-saving methods through the Rust programming language in tandem with goclassy's original Go programming language. Rust is implemented here as it is designed for the creation of more complex pipelines.

Datasets:

The Common Crawl corpus, which is a large web crawl dataset used to extract diverse knowledge, will be used for raw data. We will analyze data pipelines used to make the OSCAR dataset to filter, classify, and clean text data into a "high quality" dataset ready for NLP algorithms to train on. The OSCAR dataset is a large-scale multilingual corpus obtained by language classification and filtering of the Common Crawl corpus. This dataset is publicly available and is useful for training language models as well.
Link to Common Crawl: https://commoncrawl.org/get-started.
Link to OSCAR project: https://oscar-project.org/

Methods:

I. **Examining the Literature:** The primary focus of this project is the OSCAR dataset. Through their documentation and publications, we hope to learn and then test some of their methods in creating a high-quality dataset. [4]

II. **Asynchronous Pipeline "goclassy":** Goclassy is a pipeline used to create the OSCAR dataset in the beginning stages. It is based on a pre-made pipeline called "fastText pre-processing pipeline" which is a method to download, extract, filter, clean, and classify data extracted from Common Crawl. We will dive into the methodologies of the

fastText pipeline, modifications made on goclassy by parallelizing the pipeline in an asynchronous manner, and the results/benefits of the research. [5]

III. **An Audit of Web-Crawled Datasets:** The use of low-resource languages like “Go” for goclassy creates consequences for the quality of the resulting datasets. Evaluations on the quality of data are made on five major public datasets (CCAligned, ParaCrawl, WikiMatrix, OSCAR, mC4), and our team will look into the results and consequences of using low-resource languages.[6]

IV. **“Ungoliant” Pipeline Creation:** Our team looks into the next step of OSCAR that took the goclassy Pipeline and improved on it to create the Ungoliant pipeline. Taking a look at the new document processing methods used, we will compare the goclassy pipeline with the new Ungoliant pipeline. [7]

V. **Novel Processing of Ungoliant-retrieved Data:** There are concerns over curating datasets that remove noisy, irrelevant sentences as well as the ability of querying datasets based on desired document qualities. Our team will go over these ideas with an article that provides options for alleviating these concerns. [8]

Evaluation:

Our project aims to collect and clean textual data through the use of pipelines like goclassy and Ungoliant, confirming that the resulting data will be a key evaluation metric for our project. Our team will cross-check the results found in the articles with our own pseudo-run-through of the OSCAR dataset creation workflow. Our team will measure a way to check for deduplication to make sure the data remains non-repetitive. Additionally, our team wants to confirm the article's findings of using novel filtering and querying techniques for creating high-quality datasets.

Bibliography:

[1] Gao, Leo, et al. *The Pile: An 800 GB Dataset of Diverse Text for Language Modeling*, 31 Dec. 2020

[2] Endrédy, István, and Attila Novák. “More effective boilerplate removal - the Goldminer algorithm.” Polibits, vol. 48, 2013, pp. 79–83, https://doi.org/10.17562/pb-48-10.

[3] Penedo, Guilherme, et al. *The Refined Web Dataset for FalconLLM: Outperforming Curated Corpora with Web Data, and Web Data Only*, 1 June 2023

[4] Suarez, Pedro Ortiz. “Oscar.” *OSCAR*, 22 Feb. 2023, oscar-project.org/.

[5] Suarez, Pedro Javier Ortiz, et al. “Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures.” *Proceedings of the Workshop on Challenges in the Management of Large Corpora*, 22 July 2019, pp. 9–16, https://doi.org/https://doi.org/10.14618/ids-pub-9021.

[6] Kreutzer, J., et al. “Quality at a glance: An audit of web-crawled multilingual datasets. Transactions of the Association for Computational Linguistics”, March 2021, https://doi.org/10.1162/tacl_a_00447

[7] Abadji, Julien, et al. “Ungoliant: An Optimized Pipeline for the Generation of a Very Large-Scale Multilingual Web Corpus.” *Proceedings of the Workshop on Challenges in the Management of Large Corpora*, 12 July 2021, pp. 1–9, https://doi.org/https://doi.org/10.14618/ids-pub-10467.

[8] Abadji, Julien, et al. “Towards a Cleaner Document-Oriented Multilingual Crawled Corpus”, Jan. 2022, https://doi.org/10.48550/arxiv.2201.06642