# MultiHumES: Multilingual Humanitarian Response Dataset for Extractive Summarization

**Jenny Paola Yela-Bello**
EPFL
Lausanne, Switzerland
ypyelab@gmail.com

**Ewan Oglethorpe**
Data Friendly Space
Virginia, United States
ewan@datafriendlyspace.org

**Navid Rekabsaz**
Johannes Kepler University
Linz, Austria
navid.rekabsaz@jku.at

## Abstract

When responding to a disaster, humanitarian experts must rapidly process large amounts of secondary data sources to derive situational awareness and guide decision-making. While these documents contain valuable information, manually processing them is extremely time-consuming when an expedient response is necessary. To improve this process, effective summarization models are a valuable tool for humanitarian response experts as they provide digestible overviews of essential information in secondary data. This paper focuses on extractive summarization for the humanitarian response domain and describes and makes public a new multilingual data collection for this purpose. The collection – called MultiHumES – provides multilingual documents coupled with informative snippets that have been annotated by humanitarian analysts over the past four years. We report the performance results of a recent neural networks-based summarization model together with other baselines. We hope that the released data collection can further grow the research on multilingual extractive summarization in the humanitarian response domain.

## 1 Introduction

The disaster risk management cycle consists of four stages: mitigation, preparedness, response, and recovery (Alexander, 2002). The review of secondary data sources (i.e., reports, news, and other forms of text data) is embedded in all these stages, with varying levels of importance from stage to stage. The work of secondary data review is characterized by a high and ever-increasing amount of information to be analyzed. At the same time, typically, only a small workforce is available to analyze such information. Early in the response phase, namely in the first 72 hours after a disaster strikes, secondary data review to gain situational awareness is essential, as it brings to light which type of relief activities to undertake. After this stage, primary data collection (such as surveys) begins while still supported by the secondary data review processes.

The disaster information cycle by its part consists of the collection, collation, analysis, dissemination, decision-making, and reporting stages. An effective summarization tool can provide meaningful support for the collection stage when analysts prioritize what documents to read first (i.e., offering an overview of a document). In the collation stage, when analysts need to take and merge the most important findings from several documents, and even in the reporting stage, analysts are asked to bring in a few sentences containing the key findings of what they have written. An auto summarization system aims to provide analysts with a starting point from which they can continue their work rather than replace it. Such a system can significantly save time in the overall disaster information cycle.

In this work, we take the initial steps of creating such an extractive summarization system for humanitarian responders by curating and releasing the novel publicly available Multilingual Humanitarian Response Dataset for Extractive Summarization (MultiHumES)[1]. This collection is annotated by humanitarian experts and consists of the data related to various disasters around the globe that occurred in the last four years. Our contribution in this work occurs in two ways: the collection and consultation process for the possible release of the dataset and the dataset curation for performing extractive summarization tasks.

The dataset consists of approximately 50K documents in three languages: English, French, and Spanish. Among these documents, approximately 35K are annotated with informative snippets and can be used for the training and evaluation of ex-

[1] https://deephelp.zendesk.com/hc/en-us/sections/360011925552-MultiHumES

*Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1713–1717
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

tractive summarization models. We evaluate the performance of LEAD4, TextRank (Mihalcea and Tarau, 2004) – an unsupervised graph-based model –, and NeuSum (Qingyu et al., 2018) – a recently created supervised neural model in the dataset.

To the best of our knowledge, this is the first multilingual dataset released for summarization in the humanitarian domain. The most similar initiatives are done by (Alam et al., 2020), which released a social-media based dataset for classification in the humanitarian domain, and by Appen, which released a set of short messages from social media and news articles for classification in the humanitarian domain [2].

This article is structured as follows: Section 2 describes the background, annotation process, curation process, and main statistics of the MultiHumES collection. Section 3 explains the experiment design of the summarization models on the dataset, and Section 4 presents and discusses the results.

## 2 MultiHumES Collection

This section first provides the background on the humanitarian response domain and the ecosystem from which the collection originated. Then it explains the annotation process, the curation process and finalizes presenting key statistics from the collection.

### 2.1 Collection Background

The collection originated from a multi-organizational platform called DEEP[3]. The DEEP platform was created due to a direct need for effective secondary data management during the Nepal 2015 earthquake response. The platform facilitates classifying primarily qualitative information with respect to analysis frameworks and allows for collaborative classification and annotation of secondary data. To date, the platform has processed almost 250k manually annotated snippets across 1.7k humanitarian projects across the globe.

This research dataset contains the documents analyzed from 2016 to 2019 related to projects that occurred within 159 countries. Approximately 46% of the documents came from media sources, 29% from international organizations, and the rest from various organizations such as United Nations agencies, governments, academic and research institutions, NGOs, donors, and Red Cross/Red Crescent Movement.

Although 82% of the uploaded documents were from publicly available sources, and more than 96% are labeled as non-confidential, we made an additional consultation process with the involved organizations to ensure that the released collection preserves the privacy and dignity of any affected populations discussed in the reports.

### 2.2 Collection Annotation Process

Taggers (or humanitarian annotators) are trained in analytical standards and thinking. While undertaking secondary data review, information is selected if it fits within a given project's scope and can lead to more appropriate decision-making in a given humanitarian crisis. Key information relevant to understanding unmet needs and their underlying factors is captured and categorized into a commonly agreed analysis framework (or taxonomy), enabling a comprehensive and holistic understanding of humanitarian needs. Detailed categories and sub-categories in line with global standards and taxonomies (geographical area assessed, sectors and sub-sectors, demographic and specific needs groups, etc.) are also labeled. These selected snippets of text fill critical analysis information needs that are essential for strategic and programmatic decision-making in humanitarian response.

One of the major features of the collection is that the data has gone through a rigorous quality control process to ensure standardization, accuracy, and comprehensiveness. Designated quality control experts undertake this process in addition to peer-review and continuous training. The efficacy of the quality control efforts is seen in an external quality assessment, which shows an inter indexer consistency metric of 0.97[4].

### 2.3 Collection Curation

The original dataset consists of the plain text of 66,412 documents, the snippets extracted from each document by humanitarian analysts, and in some cases, the humanitarian classification label related to each snippet. We parsed the documents provided in HTML format to extract their text content. We also filtered out documents with 50 tokens

[2] https://appen.com/datasets/combined-disaster-response-data/

[3] https://thedeep.io/

[4] https://github.com/IFRCGo/Tagging-QAQC

or less. As verified manually, these documents were either title of web pages or small descriptions of attached documents.

The remaining documents are in a variety of languages, but mainly in English, French or Spanish. For automatically identifying the languages, we used the `langdetect` library. We kept documents in the three mentioned languages, resulting in 50,380 documents.

Note that a highlight or relevant text snippet is a sentence or a set of sentences that a humanitarian analyst deems as providing relevant information on a specific category embedded in a conceptual framework. We follow the assumption that these text snippets are not only relevant for a specific analyst and for a specific topic but also for the general summary of the document. Therefore, we combine all text snippets of a document into one document summary, which is used as the ground truth for the extractive summarization task.

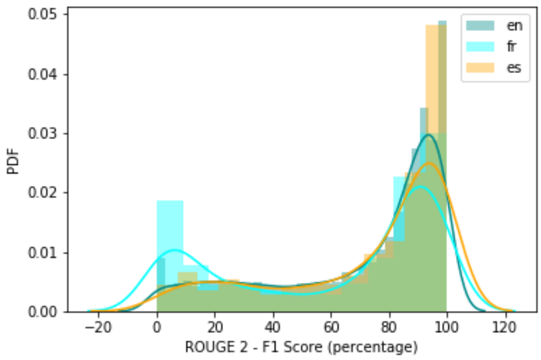

Figure 1: Oracle performance across languages.

Following Qingyu et al. (2018), we constructed an Oracle version of those summaries. Figure 1 shows the ROUGE 2 F1 Score behavior of the Oracle summaries across languages. The x-axis shows the ROUGE 2 F1 Score in a percentage format, and the y-axis shows the probability density function for the kernel density estimation. Based on these results, we see that even with an average score of 70

We evaluated some of the documents in which that score was 0 and found that some were summaries in a language different from the source document. Some had extra characters within the text caused by problems reading the texts from PDF files and implied differences between the document and summary sentences.

After conducting these preprocessing steps, 35,567 documents remained used to create the extractive summarization collection.

### 2.4 Collection Statistics

| | **English** | **French** | **Spanish** |
|---|---|---|---|
| # of documents | 29351 | 4311 | 1904 |
| # of tokens | 222918 | 58921 | 49609 |
| Median number of sent in doc | 30 | 16 | 23 |
| Median number of sent in sum | 5 | 4 | 5 |

Table 1: Dataset Statistics

Table 1 reports the total number of documents, unique tokens, and the median values of the number of sentences per document and summary. The distributions of the number of sentences per document and summary are highly skewed, such that 80% of the documents have less than 191 sentences, and 80% of the summaries have less than 11 sentences.

## 3 Experiment Design

This section explains the baseline extractive summarization models and their corresponding parameter settings.

### 3.1 Baseline Models

**LEAD4** LEAD-$n$ is an algorithm that selects the first $n$ sentences from a document as its summary. It is a simple but strong baseline for extractive summarization models, created based on the assumption that the first sentences in a document are the most informative ones. In our experiments, we use $n = 4$ as it shows the best performance on the validation set.

**Text Rank** Text Rank is a graph-based model that ranks text units from most relevant to least relevant by using text units as vertices and the similarity between text units as edges. Given this ranking and a fixed length of the desired output summary, the model produces a summary in an unsupervised manner (Mihalcea and Tarau, 2004).

**NeuSum** NeuSum is a neural extractive summarization method that employs a hierarchical document encoder to produce sentence representations. It also uses a sentence extractor to iteratively extract sentences from sentence representation extracts according to their overall contribution to the

|  | ROUGE-1 F1 |  |  | ROUGE-2 F1 |  |  |
|---|---|---|---|---|---|---|
| **Model** | **en** | **es** | **fr** | **en** | **es** | **fr** |
| *Oracle* | 0.713 | 0.761 | 0.668 | 0.706 | 0.733 | 0.619 |
| LEAD4 | 0.389 | 0.454 | 0.438 | 0.318 | 0.378 | 0.344 |
| TextRank | 0.419 | 0.456 | 0.418 | 0.289 | 0.320 | 0.277 |
| NeuSum | **0.474** | **0.531** | **0.470** | **0.380** | **0.424** | **0.358** |

Table 2: Models' ROUGE scores by language.

performance of the current summary (Qingyu et al., 2018).

The hierarchical document encoder is composed of two Recurrent Neural Networks (RNNs). The first network encodes each word in a sentence (sentence encoder), while the second network encodes each sentence given its context in the document (document encoder). Bidirectional Gated Recurrent Units (BiGRU) (Cho et al., 2014) are used for both RNNs.

### 3.2 Training and Evaluation

We partition the data into training, validation, and test sets for each language with portions of 70%, 10%, and 20%, respectively.

To evaluate the models, we used ROUGE-1 F1 and ROUGE-2 F1 as standard metrics to evaluate extractive summarization tasks. For each one of the models in each language, the 95% confidence interval of the measures' average values was calculated. The results are reported as significantly different if there is no intersection between the intervals.

For the ROUGE-1 F1 metric, we found that NeuSum performed statistically better for all languages. For the English corpus, TextRank performed statistically better than the LEAD4 baseline. For the ROUGE-2 F1 metric, we found that all results per language were significantly different.

### 3.3 Parameter Setting

We used the TextRank model implemented by the gensim library (Řehůřek and Sojka, 2010). This implementation uses the Okapi Best Matching 25 similarity measure (BM25+) to measure the similarity between sentences (Barrios et al., 2016). We modified the preprocessing steps and fixed the summaries' length to 100 words to have coherent results with the TextRank model.

For the NeuSum model, we used the word2vec SkipGram model (Mikolov et al., 2013) created on a corpus of Wikipedia provided by the gensim library (Řehůřek and Sojka, 2010). This provided embeddings trained in a similar corpus with a similar dimensionality for the three languages considered. We used 100 as the embedding dimension and 100,000 as the vocabulary size. The pre-trained embeddings covered 51.85% of the English vocabulary, 60.54% of the French vocabulary, and 70.88% of the Spanish vocabulary.

We set 200 sentences as the maximum number of sentences per document and 80 tokens as the maximum length of a sentence. The length of the output summary was set to 4. We ran the model for 40 epochs with a batch size of 64. The rest of the parameters were set as proposed in the original paper.

## 4 Results and Discussion

Table 2 reports the performance on the test sets. It can be seen that the best performing model was NeuSum, which improved by up to 10 points compared to the TextRank and LEAD4 performance.

The French corpus had the lowest performance, but the fact that the Oracle presents a low performance indicates that this result may be more related to the corpus's nature or the preprocessing of the data rather than the model.

LEAD4 bases its success on the position of the relevant sentences. TextRank bases its success on the content of the relevant sentences. NeuSum higher performance may be explained by the joint encoding of the position information and the content information of a sentence.

It is also key to remember that we truncated each document to a maximum length of 200 sentences because of our GPU capacity. This meant that reports such as the *Humanitarian Needs Overview* with around 60 pages (900 sentences) were reduced to 200 sentences and were therefore not well summarized. Our neural approach for the summarization task of humanitarian documents was useful for setting a precedent of how good these models are in the domain. However, until higher computational resources are more readily available, it would be necessary to have a simpler model to treat longer documents.

## 5 Conclusions and Future Work

Automatic summarization in the humanitarian domain is crucial for supporting fast and effective responses to crises. To facilitate this process, we provide MultiHumES, a novel multilingual collection for extractive summarization in the humanitarian response domain. The collection enables the training and evaluation of machine/deep learning-based models, revealing new horizons for research in this domain. The collection consists of approximately 50k documents, from which 35K have related annotated snippets by experts. We test and evaluate the performance of three strong baselines on the collection.

We consider the following points as potential future directions. First, a key aspect of automatic summarization for the humanitarian response domain is the human evaluation of the output. It is indeed important to understand whether the reported performance results correlate with the evaluation of domain experts. The second direction is learning multilingual models, trained over all the data, and investigating whether this approach can improve performance, especially in languages with a smaller amount of available training data.

## Acknowledgements

We would like to thank the Data Friendly Space team, and the DEEP board for their support. Jenny Paola Yela-Bello thanks the International Federation of Red Cross and Red Crescent societies for supporting her research (under contract CLMX021702).

## References

F. Alam, H. Sajjad, M. Imran, and F. Ofli. 2020. Standardizing and benchmarking crisis-related social media datasets for humanitarian information processing. *arxiv, 2020.*

David Alexander. 2002. *Principles of Emergency Planning and Management*. Terra publishing.

Federico Barrios, Federico Lopez, Luis Argerich, and Rosita Wachenchauzer. 2016. Variations of the similarity function of textrank for automated summarization. *Argentine Symposium on Artificial Intelligence (ASAI) 2015 - 44 JAIIO (September 2015).*

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, page 404–411.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems.*

Zhou Qingyu, Yang Nan, Wei Furu, Huan Shaohan, Zhou Ming, and Zhao Tiejun. 2018. Neural document summarization by jointly learning to score and select sentences. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 654–663.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.