GOUD.MA: A NEWS ARTICLE DATASET FOR SUMMARIZATION IN MOROCCAN DARIJA

Abderrahmane Issam
Archipel Cognitive
Casablanca, Morocco
abderrahmane.issam@outlook.com

Khalil Mrini
University of California San Diego
La Jolla, CA, USA
khalil@ucsd.edu

ABSTRACT

Moroccan Darija is a vernacular spoken by over 30 million people primarily in Morocco. Despite a high number of speakers, it remains a low-resource language. In this paper, we introduce GOUD.MA: a dataset of over 158k news articles for automatic summarization in code-switched Moroccan Darija. We analyze the dataset and find that it requires a high level of abstractive reasoning. We fine-tune the Arabic-language BERT (AraBERT), and the language models for the Moroccan (DarijaBERT), and Algerian (DziriBERT) national vernaculars for summarization on GOUD.MA. The results show that GOUD.MA is a challenging summarization benchmark dataset. We release our dataset publicly in an effort to encourage the diversity of evaluation tasks to improve language modeling in Moroccan Darija.[1]

1 INTRODUCTION

Moroccan Darija is the form of Arabic spoken in Morocco. According to the 2014 census, 91% of people in Morocco – over 30 million people – speak Moroccan Darija. However, Moroccan Darija remains a low-resource language with very few existing text datasets (Voss et al., 2014; Mrini & Bond, 2017; Outchakoucht & Es-Samaali, 2021), similarly to Arabic dialects and African vernaculars.

The emergence of language models, most notably BERT Devlin et al. (2019), has pushed the state of the art in various NLP tasks, mostly in English and other high-resource languages. There has been growing interest in offering the same level of language understanding performance and NLP services to speakers of North African dialects of Arabic. Tunisia's TunBERT (Messaoudi et al., 2021) is an R&D collaborative effort. Abdaoui et al. (2021) introduce DziriBERT for Algerian Darja. DarijaBERT is the language model trained for Moroccan Darija[2]. All three language models are trained on social media data, amid a common scarcity in training datasets from other domains.

In this paper, we introduce GOUD.MA: a dataset of 158k news articles for abstractive summarization in code-switched Moroccan Darija. To the best of our knowledge, our dataset is the first summarization benchmark in Moroccan Darija. We analyze our dataset and show that it requires a high level of abstractive reasoning compared to a popular summarization benchmark dataset in English. We propose to train two summarization heuristics and train three language models on this task. The results show that the language models have not yet achieved a high level of abstractive reasoning over the GOUD.MA dataset, and that this dataset offers a challenging summarization problem. We open-source this dataset in an effort to encourage the diversity of evaluation tasks for language modeling in Moroccan Darija.

2 RELATED WORK

Moroccan News Article Datasets. Previous work has proposed news article datasets from Moroccan sources, but in Modern Standard Arabic and not in Moroccan Darija.

[1] Our Github repository: https://github.com/issam9/goud-summarization-dataset

[2] Information about DarijaBERT is available here: https://github.com/AIOXLABS/DBert

Jbene et al. (2021) propose the Moroccan News Article Dataset (MNAD): a dataset of 418k news articles. The authors collect the articles from four major Moroccan online news sources: Akhbarona.ma, Hespress.ma, Hibapress.com, and Le360.ma. They also collect 19 different categories for the news articles, and propose to perform text categorization on their dataset.

Boukil et al. (2018) propose a CNN-based model for text categorization for news articles from three Moroccan online sources: Hespress, Akhbarona, and Assabah. The total number of articles they collect is 112k. Likewise, these news articles are in Arabic and not in Moroccan Darija.

Moroccan Darija Datasets. There is a scarcity of text datasets in Moroccan Darija. These datasets can be divided into two categories: word-level bilingual resources, and free text datasets.

Many word-level bilingual resources for Moroccan Darija focus on word-level and context-free translation. Tachicart et al. (2014) introduce the Moroccan Dialect Electronic Dictionary (MDED): a word-level Moroccan Darija-Standard Arabic bilingual dictionary. Mrini & Bond (2017) build the Moroccan Darija WordNet (MDW): a Moroccan Darija extension of the Open Multilingual WordNet (Bond & Foster, 2013). Both the MDW and MDED were used to compute similarities of single words in Moroccan Darija to their translations in Modern Standard Arabic (Tachicart et al., 2014) and other languages (Mrini & Bond, 2018). Outchakoucht & Es-Samaali (2021) propose DODa: the Darija Open Dataset, a collaborative dataset of words in Darija and their equivalent in English.

Free text datasets in Moroccan Darija are more frequently used for deep learning applications. Tratz et al. (2013) present an annotated corpus of mixed-script and code-switched Moroccan Darija tweets. Samih & Maier (2016) study code-switching detection in Moroccan Darija using their own dataset. Mihi et al. (2020) introduce a dataset for sentiment analysis over tweets in Moroccan Darija. There has been significant effort in Arabic dialect identification involving text data in Moroccan Darija (Nour-Eddine & Abdelkader, 2015; Tachicart et al., 2017; Abdelali et al., 2020; Issa et al., 2021). Moroccan Darija has recently obtained its own Wikipedia, where the standardization of writing has been a hot topic for debate (Sedrati et al., 2020). It is not yet feasible to train an entire language model on the Moroccan Darija Wikipedia alone, as it remains relatively small.

3 Dataset Construction

3.1 Goud.ma

Goud.ma[3] is a Moroccan news website, created in 2011 by Ahmed Najim and edited by the Goud Media company. Goud.ma is a free, ad-supported online media. "*Goud*" means "straight-forward" in Moroccan Darija.

All articles on Goud.ma are written in the Arabic script. All headlines are in Moroccan Darija, whereas articles may be in Moroccan Darija, in Modern Standard Arabic, or a mix of both (code-switched Moroccan Darija). This generally represents the use of Moroccan Darija when written in the Arabic script, as code-switching with Modern Standard Arabic is frequent.

3.2 Obtaining Goud.ma Articles

The Goud.ma website is built on top of Wordpress, therefore we rely on their .json Rest API to access their data, which is retrieved in .json format from each page. A single page contains 10 articles and their headlines, and other information such as the link to the article on the Goud.ma website and the date it was published. We loop through these pages and get the articles, headlines and article categories. Since the articles come with HTML tags and with characters that appear in their Unicode encodings, we rely on BeautifulSoup to parse the HTML and extract only the content of the HTML paragraph tag (`<p>`). Most of articles on Goud.ma start with the name of the author and sometimes the city where the event that is being described in the article has happened, thus, we remove this part. We also removed website links, and some HTML tags that were still in the articles. Some articles and headlines are very short, we only keep entries where the number of words in the article and the headline are higher than 30 and 3 respectively. The only cleaning we needed to apply to the headlines is parsing them using BeautifulSoup to decode the Unicode encodings back to their string format.

[3] https://www.goud.ma/

	Articles	**Headlines**
The number of tokens	26,780,273	2,143,493
The number of unique tokens	1,229,993	236,593
Minimum number of tokens	32	4
Maximum number of tokens	6,025	74
Average number of tokens	169.19	13.54

Table 1: GOUD.MA Dataset statistics.

Category	**Category translation**	**Number of articles**
الرئيسية	Main	104,724
آش واقع	What's happening	98,569
تبركيك	Gossip	16,867
كود سبور	Goud Sport	13,236
آراء	Opinions	8,239
ميديا وثقافة	Media and Culture	7,579
كود تيفي	Goud TV	6,966
الزين والحداكة	Beauty and Sharpness	5,223
جورنالات بلادي	National Newspapers	4,549
كود	Goud	1

Table 2: Distribution of categories over articles of the GOUD.MA dataset.

4 Dataset Analysis and Statistics

4.1 Overview

After cleaning the dataset, we end up with 158,282 articles and their headlines. Table 1 shows simple statistical features of the dataset, some features are directly affected by the cleaning we applied, like the minimum number of words. We split our dataset randomly into 88% training, 6% validation and 6% test.

One article can be tagged with up to four categories on GOUD.MA, but most articles have two categories. Most of the time, articles are tagged as الرئيسية (Main page) and another category. Table 2 shows the number of articles of each category and its translation to English. The diversity of these categories shows the diversity of topics covered in the GOUD.MA website.

4.2 Summary Analysis

Summarization datasets differ in their levels of abstractiveness and extractiveness. Grusky et al. (2018) define two measures to quantify the extractiveness of summaries: extractive fragment coverage and extractive fragment density. Given an article $A = \langle a_1, a_2, ..., a_n \rangle$ and its summary $S = \langle s_1, s_2, ..., s_n \rangle$, the fragments $F(A, S)$ are the set of tokens that are shared between A and S identified greedily. Coverage is the percentage of summary words that are part of the fragments, while density is defined as the average length of the fragment of each word in the summary. More formally, we compute coverage and density as follows:

$$\text{Coverage}(A, S) = \frac{1}{|S|} \sum_{f \in F(A,S)} |f| \tag{1}$$

$$\text{Density}(A, S) = \frac{1}{|S|} \sum_{f \in F(A,S)} |f|^2 \tag{2}$$

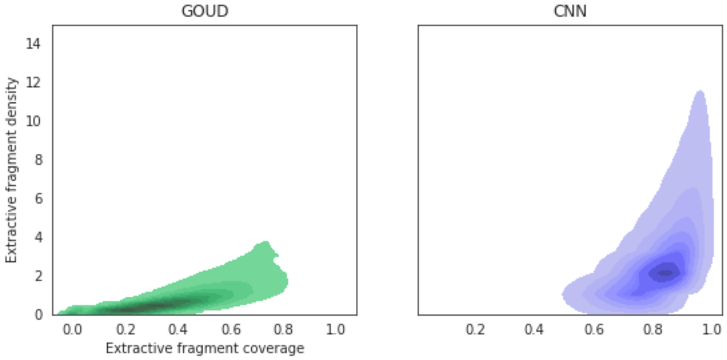

Figure 1: Coverage and density of our GOUD.MA dataset (left) compared to the CNN-Daily Mail summarization benchmark dataset (Hermann et al., 2015) (right).

In Figure 1, we show the distribution of coverage and density of our dataset in comparison to the CNN-Daily Mail summarization benchmark dataset. The CNN-Daily Mail dataset was introduced as a question answering dataset (Hermann et al., 2015), but it is currently widely used as a summarization dataset (See et al., 2017), where article highlights that existed as bullet points are concatenated to form summaries. From Figure 1 and as was noted in Grusky et al. (2018), CNN-Daily Mail is skewed towards extractiveness, and in comparison, our GOUD.MA dataset is more on the abstractive side, as it is characterized by low coverage and density.

5 EXPERIMENTS

In this section, we show how one can use the GOUD.MA dataset for abstractive summarization. We propose the following task definition: we train our model to summarize news articles, with the headlines as the target text. The goal of our experiments is to establish baselines for Moroccan Darija summarization on the GOUD.MA dataset, and thus encourage further development of summarization systems in this low-resource language.

5.1 SETUP

To measure the performance of our models, we use the F1 variant of ROUGE (Recall-Oriented Understudy for Gisting Evaluation), which is a popular metric for text summarization. In our case, we only consider ROUGE-1, ROUGE-2 and ROUGE-L variants. ROUGE-1 measures the overlap of unigrams between the reference summary and the candidate, while ROUGE-2 considers bigrams instead. We also compute ROUGE-L metric which computes the longest common subsequence. Unlike ROUGE-n, ROUGE-L takes into consideration the order of words in the sentence, and doesn't require a predefined n-gram length.

5.2 METHODS

We propose to use two widely used summarization heuristics and three deep learning methods based on fine-tuned language models.

The two summarization heuristics are:

- **Extractive Oracle Fragments:** This is the result of concatenating the fragments in $F(A, S)$ (see §4) in the order they appear in the summary, and represents the best possible performance of an extractive system.
- **Lead-1:** This strategy simply considers the first sentence of an article as a candidate summary. Since the first few sentences introduce the content of an article, this can be a competitive baseline.

Methods	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Extractive Oracle Fragments	50.20	23.91	50.20
Lead-1	14.52	5.08	13.44
AraBERT	23.08	8.98	22.06
DarijaBERT	19.41	6.64	18.48
DziriBERT	17.98	5.83	17.22

Table 3: Summarization results on the test set of the GOUD.MA dataset.

We fine-tune three language models for this abstractive summarization task:

- **AraBERT[4]:** Antoun et al. (2020) introduce a BERT model (Devlin et al., 2019) trained for the Arabic language. AraBERT was trained on the Arabic-language Wikipedia, and on two additional Arabic-language news corpora of a cumulative size of 2.5 billion tokens.
- **DarijaBERT[5]:** this language model was developed and made available by AIOX Labs, an AI company based in Rabat, Morocco. DarijaBERT is a language model for Moroccan Darija written in the Arabic script. It is pre-trained on stories written in Darija, comments from 40 Moroccan Youtube channels, and tweets crawled using a keyword list. No paper was published detailing the training setup.
- **DziriBERT[6]:** Abdaoui et al. (2021) introduce a language model based on BERT for Algerian Darja. We use this model as Algeria's national vernacular is largely mutually intelligible with Moroccan Darija. DziriBERT is trained on 1.2 million tweets emanating from major Algerian cities, but the authors consider this to be on the smaller side of language modeling training dataset sizes.

All three language models are available on Hugging Face (Wolf et al., 2020).

5.3 TRAINING DETAILS

For the language models, we build encoder-decoder models that are warm-started by leveraging pre-trained BERT checkpoints, this was introduced in Rothe et al. (2019), and was shown to improve the results in multiple tasks including abstractive summarization. While in Rothe et al. (2019) they investigate multiple combinations, we use the BERT setup that achieved the best results in abstractive summarization, which is to warm-start both the encoder and the decoder, and tie their weights, this also helps reduce the memory footprint.

While BERT Models can handle a maximum of 512 tokens, we truncate the articles to a maximum length of 256, and the headlines to 32.

We train our models for 20 epochs, with a per device batch size of 8, and 4 gradient accumulation steps, this results in a total batch size of 128. We specify a weight decay of 0.01 and set the number of warmup steps to 1000. Half-precision was used to speed up the training.

During text generation, we use Beam Search with 5 as the number of beams, we enable early stopping and prevent the model from repeating bigrams. The minimum and maximum length of the sequence to be generated was set to 4 and 32 respectively to match the length distribution of the summaries in our dataset.

5.4 RESULTS AND DISCUSSION

We show the test results of our summarization experiments on the GOUD.MA dataset in Table 3.

The results of the language models are fairly close to the Lead-1 baseline: this shows that this abstractive summarization task is difficult for the existing language models for Modern Standard Arabic and North African Arabic dialects. We notice that AraBERT performs the best, and that DarijaBERT and DziriBERT are close to each other. A number of factors can explain this performance.

[4] AraBERT is available here: `https://github.com/aub-mind/araBERT`

[5] DarijaBERT is available here: `https://github.com/AIOXLABS/DBert`

[6] DziriBERT is available here `https://github.com/alger-ia/dziribert`

Article	لونصات شركتي أفريقيا وأفريقيا غاز خدمة الأداء بالتيليفون، اللي غادي تنطلق ابتداء من الاثنين ١٥ يونيو، وغادي يتم تعميم الخدمة والعمل بها بجميع شبكاتها وفروعها. وحسب بلاغ للشركة، توصلات بيه كود، فهذ الخدمة يمكن الاستفادة منها فجميع ليسطاسيون إفريقيا، ومحطات “Auto Go” وقهاوي “Oasis Cafe” ومحلات ميني براهيم ، بالإضافة لموزعي الكَاز البقالة ومحطات ألو كَاز، باش يلبيو الحاجة المتزايدة على هذ الخدمة بشكل مزيان فالسوق الوطنية. وغادي تعطي طريقة الدفع الجديدة للزبناء إمكانية دفع ثمن مشترياتهم بأمان وسهولة من تيليفوناتهم، وهذ الحل كيوفر مجموعة من المزايا كتتوافق مع حالة الطوارئ الصحية اللي كتعرفها البلاد بسبب فيروس كورونا المستجد. غيحتاج الزبون توفير محفظة رقمية “MWallet” من عند الفاعل البنكي، أو إحدى مؤسسات الدفع عبر الهاتف، باش يتم الدفع على QR Code المسجل في جهاز الدفع الإلكتروني.
Reference summary	فأفريقيا يمكن ليك دابا تاكل الطواجن وتعمر الگازوال وتشري من ميني براهيم غير بالتلفون
AraBERT summary	إفريقيا غاز و أفريكول لونصات خدمة الأداء بالتيليفون
DarijaBERT summary	لمحاربة كورونا . . شركة افريقيا لونصات خدمة الاداء بالتيليفون
DziriBERT summary	لونكيط لونصا افريقيا لونصات خدمة لاسيرونس بالتيليفون . . وها التفاصيل

Table 4: An example news article from the GOUD.MA test set, along with the corresponding reference summary (headline) and generated summaries from AraBERT, DarijaBERT, and DziriBERT.

First, AraBERT has the largest training dataset size of the three models. Second, many articles contain Modern Standard Arabic. Finally, AraBERT is the only language model in our list to be trained on news corpora, whereas DarijaBERT and DziriBERT are trained on social media text data.

The scores of the language models are well below the score of the Extractive Oracle Fragments. This means that they are not yet able to perform abstractive reasoning on the GOUD.MA dataset. Therefore our dataset presents a reasonable challenge for abstractive summarization in Moroccan Darija. The GOUD.MA dataset can serve to evaluate improvements in North African language models in yet another task.

We show an example of a news article from GOUD.MA in Table 4, along with the headline (reference summary) and the generated summaries from all three language models. This news article is mainly written in Moroccan Darija, with a few code-switched expressions in French and Modern Standard Arabic. Whereas the reference summary is abstractive, the generated summaries are more extractive. We notice AraBERT has learned words from Moroccan Darija (لونصات), whereas DziriBERT picks up on click-bait patterns (وها التفاصيل – "*and here are the details*"). DarijaBERT has no repetitions (لونكيط لونصا) or hallucinations (أفريكول) unlike the other two language models.

6 CONCLUSIONS

In this paper, we introduce GOUD.MA: a dataset of over 158k news articles for abstractive summarization in code-switched Moroccan Darija. To the best of our knowledge, it is the first summarization benchmark for Morocco's national vernacular. We show that summarization on our dataset requires a high level of abstractive reasoning compared to CNN-Daily Mail, a popular benchmark for summarization in English. We train three language models for our proposed task. The results show that GOUD.MA is a challenging summarization dataset, and that it can serve as an additional axis of evaluation to further improve language modeling for Moroccan Darija.

REFERENCES

Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. Dziribert: a pre-trained language model for the algerian dialect. *arXiv preprint arXiv:2109.12346*, 2021.

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. Arabic dialect identification in the wild. *arXiv preprint arXiv:2005.06557*, 2020.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pp. 9–15, 2020.

Francis Bond and Ryan Foster. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1352–1362, 2013.

Samir Boukil, Mohamed Biniz, Fatiha El Adnani, Loubna Cherrat, and Abd Elmajid El Moutaouakkil. Arabic text classification using deep learning technics. *International Journal of Grid and Distributed Computing*, 11(9):103–114, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. 04 2018.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pp. 1693–1701, 2015.

Elsayed Issa, Mohammed AlShakhori, Reda Al-Bahrani, and Gus Hahn-Powell. Country-level arabic dialect identification using rnns with and without linguistic features. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pp. 276–281, 2021.

Mourad Jbene, Smail Tigani, Rachid Saadane, and Abdellah Chehri. A moroccan news articles dataset (mnad) for arabic text categorization. In *2021 International Conference on Decision Aid Sciences and Application (DASA)*, pp. 350–353. IEEE, 2021.

Abir Messaoudi, Ahmed Cheikhrouhou, Hatem Haddad, Nourchene Ferchichi, Moez BenHajhmida, Abir Korched, Malek Naski, Faten Ghriss, and Amine Kerkeni. Tunbert: Pretrained contextualized text representation for tunisian dialect. *arXiv preprint arXiv:2111.13138*, 2021.

Soukaina Mihi, B Ait, I El, Sara Arezki, and Nabil Laachfoubi. Mstd: moroccan sentiment twitter dataset. *Int. J. Adv. Comput. Sci. Appl*, 11(10):363–372, 2020.

Khalil Mrini and Francis Bond. Building the moroccan darija wordnet (mdw) using bilingual resources. In *Proceedings of the International Conference on Natural Language, Signal and Speech Processing (ICNLSSP)*, number CONF, 2017.

Khalil Mrini and Francis Bond. Putting figures on influences on moroccan darija from arabic, french and spanish using the wordnet. In *Proceedings of the 9th Global Wordnet Conference*, pp. 372–377, 2018.

Lachachi Nour-Eddine and Adla Abdelkader. Gmm-based maghreb dialect identification system. *Journal of Information Processing Systems*, 11(1):22–38, 2015.

Aissam Outchakoucht and Hamza Es-Samaali. Moroccan dialect-darija-open dataset. *arXiv preprint arXiv:2103.09687*, 2021.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. Leveraging pre-trained checkpoints for sequence generation tasks. 07 2019.

Younes Samih and Wolfgang Maier. Detecting code-switching in moroccan arabic social media. *SocialNLP@ IJCAI-2016, New York*, 2016.

Anass Sedrati, Abderrahman Ait Ali, et al. Moroccan darija in online creation communities: Example of wikipedia. 2020.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368, 2017. URL `http://arxiv.org/abs/1704.04368`.

Ridouane Tachicart, Karim Bouzoubaa, and Hamid Jaafar. Building a moroccan dialect electronic dictionary (mded). In *5th International Conference on Arabic Language Processing*, pp. 216–221, 2014.

Ridouane Tachicart, Karim Bouzoubaa, Si Lhoussaine Aouragh, and Hamid Jaafa. Automatic identification of moroccan colloquial arabic. In *International Conference on Arabic Language Processing*, pp. 201–214. Springer, 2017.

Stephen Tratz, Douglas Briesch, Jamal Laoudi, and Clare Voss. Tweet conversation annotation tool with a focus on an arabic dialect, moroccan darija. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 135–139, 2013.

Clare Voss, Stephen Tratz, Jamal Laoudi, and Douglas Briesch. Finding romanized arabic dialect in code-mixed tweets. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 2249–2253, 2014.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.