LANS: Large-scale Arabic News Summarization Corpus

Abdulaziz Alhamadani
Virginia Tech
Falls Church, VA, USA
hamdani@vt.edu

Xuchao Zhang
Microsoft,
Redmond, WA, USA
xuchaozhang@microsoft.com

Jianfeng He*
Virginia Tech
Falls Church, VA, USA
jianfenghe@vt.edu

Aadyant Khatri
Virginia Tech
Falls Church, VA, USA
aadyant@vt.edu

Chang-Tien Lu
Virginia Tech
Falls Church, VA, USA
ctlu@vt.edu

Abstract

Text summarization has been intensively studied in many languages, and some languages have reached advanced stages. Yet, Arabic Text Summarization (ATS) is still in its developing stages. Existing ATS datasets are either small or lack diversity. We build, LANS, a large-scale and diverse dataset for Arabic Text Summarization task. LANS offers 8.4 million articles and their summaries extracted from newspapers websites' metadata between 1999 and 2019. The high-quality and diverse summaries are written by journalists from 22 major Arab newspapers, and include an eclectic mix of at least more than 7 topics from each source. We conduct an intrinsic evaluation on LANS by both automatic and human evaluations. Human evaluation of 1,000 random samples reports 95.4% accuracy for our collected summaries, and automatic evaluation quantifies the diversity and abstractness of the summaries.

1 Introduction

Every day there is an abundant amount of text published on the internet, such as news articles, scientific papers, product reviews, and blogs. Therefore, the need for text summarization is compelling to make use of this information overload. For a summarized text, a good one should be concise and include the main information of the original text (Radev et al., 2002). For some languages like English, the field has developed rapidly and achieved competitive results(Zhang et al., 2020; Lewis et al., 2019; Dou et al., 2020). Unlike English, the field in Arabic has been slowly and fairly developing in the past few years; thus, it has not reached its advanced shape. In the field of Arabic Text Summarization (ATS) (Belkebir and Guessoum, 2015; AL-Khawaldeh and Samawi, 2015; Fejer and Omar, 2014; Abu Nada et al., 2020; El-Kassas et al., 2021), the dearth of a diverse and large summarization dataset is one of the main existing difficulties that ATS researchers encounter (Al-Saleh and Menai, 2016; Elsaid et al., 2022).

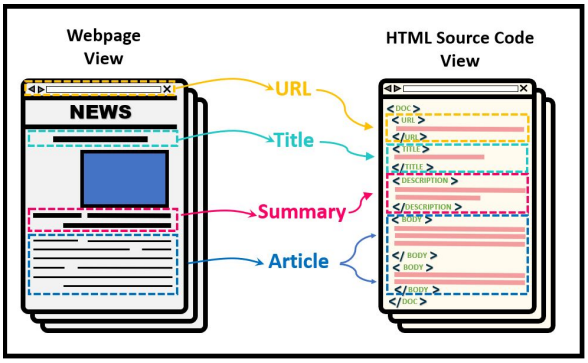


Figure 1: The Webpage view (left) shows a typical news article view. The summaries are extracted from the HTML source code view's (right) metadata (*og:description*).

Concerted efforts have been made to overcome those challenges by building various Arabic datasets for the task such that EASC (El-Haj et al., 2010), Kalimat(El-Haj and Koulali, 2013), TAC2011 (El-Ghannam and El-Shishtawy, 2014), ANT(Chouigui et al., 2021), and XL-Sum (Hasan et al., 2021), but those datasets have limitations in terms of diversity or size. Therefore, the demand for a diverse and large-scale dataset is crucial to advance the ATS field. The diversity in the ATS dataset is in twofold. The first kind of diversity exists in the Modern Standard Arabic (MSA). Even though 22 countries use MSA as an official standard language, each country has its own dialects (Dialectal Arabic) for communication. Each country's dialects have some effects on the MSA style of writing and the choice of words. For example, in a sentence describing the rounds of a soccer match, Moroccan MSA would use the word "أطوار" for "rounds" and "المواجهة" for the word "the match" while Saudi MSA would use "أشواط" and "المباراة"

*Corresponding author.

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 89–100

Corpus	# of documents	MSA Diversity	Category Diversity	Human Evaluation
EASC	153	×	×	✓
KALIMAT	20,291	×	✓	×
ANT	31,798	×	✓	×
XL-Sum	40,327	×	✓	250
LANS	> 8 millions	✓	✓	1,000

Table 1: Arabic Text Summarization Datasets comparison

respectively. Second, there is diversity in news categories. Each newspaper has different news topics, such as finance, politics, sports, health, local, international news, and more. Not all ATS datasets include both diversity aspects in one dataset (see Table 1). Thus, it is essential to build a dataset that considers both types of diversity.

In terms of size, the available ATS datasets contain a range of 100 to 41,000 training samples, which make them too small to fully train a summarization model. The performance in summarization models evidently relies on a substantial amount of applicable training samples (Völske et al., 2017; Grusky et al., 2018; Zhang et al., 2020; Lewis et al., 2019; Dou et al., 2020). Thus, we expect a large-scale dataset which is provided in this work.

To overcome the current limitations in diversity and size, we introduce a new ATS dataset (**LANS**) that includes both types of diversity and large-scale to present new opportunities to ATS models and improve their summary accuracies. To achieve MSA diversity, that is the variety of each Arab country dialects' impact on its MSA, LANS encompasses 19 Arab countries and collected articles along with their summaries of 22 popular newspapers (see Table 2). For the diversity of text categories, we consider all available news categories of each source in our ATS dataset. Thus, LANS ensures both types of diversity of MSA among the Arab countries. To overcome the size limitation, LANS provides more than 8 million news articles along with their summaries. LANS's substantial amount of articles and their summaries, plus the diversity in MSA sources and categories make it a worthy resource for ATS models.

LANS exploited the metadata of newspapers' archives to extract and build the dataset. In Figure. 1, a high-level example is shown to demonstrate where the collected information originated from two parallel views: the webpage view and its HTML source code view. The webpage view shows what a reader sees when reading a news article: the URL, title, bold part or the abstract sentence/s, and article bodies. LANS pursues the metadata attributes from the HTML source code - specifically (*og:description*) to extract the summaries from. In the webpage view, the summaries lie either in bold text or before the article's paragraphs. In the HTML source code view, the summaries lie in the metadata attributes, in our case between (*og:description*) tags, which we extracted as the news articles' summaries. After the extraction, we cleaned and filtered 11M news articles to present 8.4M articles along with their summaries.

To quantify the quality of the collected summaries and examine their summarization properties, we conducted an automatic evaluation based on 3 common metrics. Moreover, we corroborated the evaluation with human evaluation of 1,000 samples to verify the accuracy of using the abstract from the HTML source code's metadata as a summary. The human evaluation verifies that using the summary available in the metadata has a 95.4% accuracy. Considering the large size of LANS, 8.4 million, LANS can benefit the ATS field, because large datasets improve NLP tasks, such as numerous training samples for pre-trained models (Zhang et al., 2020; Lewis et al., 2019). Besides, both types of diversities create opportunities for researchers to construct more accurate ATS models.

Our main contributions are as follows: (1) We curate LANS, a large-scale ATS dataset of 8.4 million Arabic news articles paired with their summaries written by journalists between 1999 to 2019. To our knowledge, it is the largest to date. (2) LANS is collected from 22 reputable Arab newspapers to achieve high quality of diversity in MSA, and for each source, there are at least 7 topics to achieve diversity in categories. (3) To quantify the intrinsic quality of LANS, a human evaluation is conducted on 1,000 random samples and verifies 95.4% accuracy of the summaries. Plus, the automatic evaluation on the whole dataset quantifies the abstractness and properties of the summaries.

ID	Newspaper	Country	From	Articles	ID	Newspaper	Country	From	Articles
1	Elkhabar	Algeria	2014	78201	12	Hespress	Morroco	2007	91357
2	Alwasat	Bahrain	2013	23860	13	Alwatan	Oman	2014	130067
3	Gate Ahram	Egypt	2016	315655	14	Alquds	Palestine	2015	88313
4	Youm7	Egypt	2008	2039818	15	Alquds-UK	Palestine	2013	349439
5	Albayan	Emirates	1999	1137188	16	Alwatan	Qatar	2016	214405
6	Almadapaper	Iraq	2009	105925	17	Aljazira	Saudi Arabia	2001	809445
7	Aldustoor	Jordan	2000	601372	18	Alryiadh	Saudi Arabia	2004	1004893
8	Annahar	Kuwait	2007	575482	19	Alsudan Alyoom	Sudan	2016	104439
9	Alakhbar	Lebanon	2006	222215	20	Zamanalwsl	Syria	2007	128785
10	WAL	Libya	2013	141898	21	Alssabah	Tunisia	2011	166137
11	Sahara Media	Mauritania	2009	11982	22	Almasdar	Yemen	2009	102608
								Total	8,443,484

Table 2: Overall statistics of the collected articles

2 Related Work (Existing Datasets)

To the best of our knowledge, Lakhas (Douzidia and Lapalme, 2004) is considered one of the early works to build an ATS model. Due to the lack of ATS datasets at that time, Douzidia et al. translated (DUC)[1] dataset, from English to Arabic for their ATS model's evaluation (Douzidia and Lapalme, 2004). The translation used machine translation at that time which was not as accurate and advanced as these days, and that had a negative impact on the results. Moreover, other ATS models built their own datasets to evaluate their models (Al-Maleh and Desouki, 2020). Consequently, researchers built Arabic ground-truth summaries over the past years, and this section mentions the major ones.

The Essex Arabic Summaries Corpus (EASC) Dataset. EASC (El-Haj et al., 2010) is an ATS dataset, where each summary is extracted from the texts by Mechanical Turk. Its text source is two Arabic newspapers (Alrai and Alwatan) and the Arabic language version of Wikipedia. As a result, it contains 153 Arabic articles and 765 summaries (5 summaries per article). In short, EASC has high-quality human-generated summaries but it is too small and lacks diversity.

Kalimat Dataset. El-Haj et al. worked on a dataset called Kalimat (El-Haj and Koulali, 2013). It has 20,291 extractive Single-document and multi-document system summaries, and includes only 6 categories. It has been collected from only one source, which is Alwatan newspaper from Oman. The single-document summaries are generated based on their model Gen-Summ which inputs the article and its first sentence, then outputs the extractive summary. The multi-document summaries were generated for each 10, 100, and 500 articles in different categories. The generated summaries also lack human evaluation of the summaries.

Arabic News Texts Corpus (ANT) and XL-Sum. ANT (Chouigui et al., 2021), and XL-Sum (Hasan et al., 2021) are the most recent works. ANT collected 31,798 documents paired with summaries using RSS feeds from 5 Arab news sources: AlArabiya, BBC, CNN, France24, and SkyNews, while XL-Sum collected 40,327 only from BBC. ANT includes 6 categories, while XL-Sum reported none. Unlike ANT, LANS utilized the HTML source code *og:description* tag to collect the summaries which is similar to (Grusky et al., 2018). ANT is evaluated on several extractive summarization methods such as LexRank, TextRank, Luhn and LSA. XL-Sum fine-tuned mT5 on their dataset and randomly sampled 500/500 development and test set respectively. Besides, they conducted human evaluation on 250 random samples. When compared to our LANS, our work collected nearly 8 million articles with summaries from 19 Arab countries local newspapers. Moreover, experts evaluated 1,000 random summaries from LANS to substantiate the validity of the summaries.

3 LANS Dataset

This section details how LANS is collected starting from the scraping process to building the dataset and how it is shaped for public use.

3.1 Data Collection

Our main goal is to improve the ATS field by collecting and building the largest and most diverse ATS dataset. We collect newspapers from 19 countries [2]. For consistency and fairness of data col-

[1] An English text summarization dataset of news paired with human summaries. https://duc.nist.gov/

[2] There are 22 Arab countries, but 3 of them: Djibouti, The Comoros Islands, and Somalia, lack Arabic data and reliable newspapers

lection, all the TV news channels' websites are excluded, like Alarabiya, Aljazeera, Arabic CNN, and Arabic BBC because they are primarily established as TV news channels. To make our data sources comprehensive and trustworthy, we collected and listed approximately all the reliable newspapers for each country. For instance, we listed 18 reputable newspapers in Saudi Arabia. After analyzing the newspapers, we then ranked them by assigning the highest priority to the newspaper with the longest publishing history.

Next, we only select the newspapers if their content passes certain criteria:

History of published articles (archive): Each newspaper's website is inspected to examine if it has a considerable historical electronic archive to reestablish the long-history versions of a newspaper. An old reputable newspaper can be given a lower rank over a modern one if the latter has a longer historical e-archive. Thus, LANS has collected data from 1999 to 2019 see Table 2.

Diversity in categories: A newspaper should contain a variety of topics or categories (at least 7), for example, local news, international news, politics, economy, religion, culture, health, sports, art, technology, and so on.

Availability of the summary in the metadata: the metadata of a document has the hidden information of an article. The summary of an article written by the author initially lies in the metadata and also can appear in bold on the webpage or ahead of the article. The availability of the summary published by the author/journalist is the major factor in selecting the newspaper. Only the newspapers with provided summaries in the metadata are selected.

The aforementioned criteria narrow down the list of the reliable newspapers, shown in Table 2. As a result, 22 popular newspapers of 19 Arab countries have been selected for the next step from the period of time between 1999 to 2019. The wide variety of the data sources can significantly benefit the diversity of the summaries.

3.1.1 Data Scraping

Since there are 22 newspaper websites to be scraped, it is necessary to customize a code for each of them. Each code identifies the patterns, the selectors, and the URLs to be scraped. The main information scraped from each news article are the following: URL, title or (headline), article, and finally the summary or (the metadata from *og:description*). An example is shown in Table 3, which shows the scraped information from an article's webpage. For reproducibility, *Scrapy* was ideal, in our case scenario, for implementing recurring and large-scale web scraping projects. Besides, *Scrapy* supports different built-in data outputs such as JSON, XML, and CSV.

3.2 Building LANS Dataset

For the collected data to be curated so it preserves a good quality for reuse and evaluation, we detail how the data is extracted, cleaned, and preprocessed.

3.2.1 Data Extraction

Among the data formats for retrieval, the most convenient format to preserve data quality is XML. The extracted data is stored in a tree structure. Each newspaper has a dataset formatted as the following: "Item" is the root node of the tree. The root has many child nodes "Items". Each "Items", a child node, holds the extracted data of a single document (a newspaper article). The child node, "Items", has 4 child nodes of its own named: Address, Title, Article, and Summary. Each child node of the parent "Items" (Address, Title, Article, and Summary) has 1 or more grandchild nodes depending on the actual values extracted from an article's webpage. The data in this stage is not considered clean nor reliable because it contains many errors that could impact the quality of LANS. Errors can be extraneous or foreign characters, empty values, HTML code, or other common text errors. Thus, we need to clean the data. Plus, to better utilize the data in the XML files, we need to preprocess the data for the evaluation process.

Data cleaning: Initially, more than 11 million articles and their metadata are scraped. The data is laboriously examined to ensure whether the extracted articles are error-free content or not, and to ensure their validity for usage. One of the main errors was the collected articles with missing content. There are some reasons for that. One of the reasons is that many articles contain only images or videos without any textual content, because they are types of news that only report pictures or videos. The other reason for missing content is mistakes from the HTML pages, or content stored under a different selector. All articles with the mentioned errors are removed. Moreover, to clean the other errors the normalization step in the preprocessing steps below is performed. In short, the removed articles may have no title, article, or valid data. Af-

Type	Scraped info
URL	http://www.alwasatnews.com/news/1196668.html
Title	بالصور... المرخ الخيرية تنظم حملة تنظيف لمقبرة القرية
Article	قام المشاركون بإزالة الأشجار والأوساخ الضارة وتقليم الأشجار، وقد شهدت الحملة مشاركة من الأهالي من جميع الفئات العمرية، بالإضافة لأعضاء مجلس إدارة الجمعية. من جانبه، قال رئيس لجنة شئون القرية والمقبرة في الجمعية مصطفى عبدالنبي إن الحملة تأتي استكمالاً لعملية التطوير شامل للمقبرة، حيث تستعد اللجنة للبدء بالمرحلة السادسة من عملية تطوير المقبرة والتي ستشمل عمل كراسي للمظلة ورصف الطريق المؤدي من المغتسل إلى المظلة ونقل خزان الماء الرئيسي من موقعه الحالي إلى الجهة الشرقية للمغتسل وإصلاح واستكمال شراء الاحتياجات، بالإضافة إلى متابعة الخطة التطويرية بالتنسيق مع إدارة الأوقاف الجعفرية، هذا وأثنى على نشاط المشتركين في الحملة، كما قدم شكره لجميع أبناء القرية لتعاونهم لإنجاح حملة تنظيف المقبرة.
Summary	نظمت لجنة شئون القرية في جمعية المرخ الخيرية الاجتماعية، تزامناً مع رأس السنة الميلادية، حملة تنظيف لمقبرة القرية تحت شعار استثمر وقتك لآخرتك، صباح أمس الأحد ١ يناير كانون الثاني ٢٠١٧

Table 3: An example of scraped information from an Article

ter removing all the unusable articles, the number has dropped from 11,115,932 to 8,443,484 articles. After this step, the data is stored in its final XML tree format.

3.2.2 Preprocessing

Even though the data is clean at this stage, it requires preprocessing for ATS evaluation process, due to the complex and rich nature of Arabic language. The steps involve normalization, segmentation, removal of stop words, and lemmatization; in that order. This stage in Arabic is the primary stage to prepare the text for processing and transform the input text into a unified representation.

The normalization step cleans the data and removes many extraneous texts. It removes extra white spaces or tabs, foreign irrelevant characters, non-letters, and diacritics. It also replaces certain Arabic characters with a certain single character to normalize the differences in characters. Normalization also removes the "Tatweel" (character stretching) (Ayedh et al., 2016). For tatweel, a word that appears in this format "تـمـديـد" is going to be replaced with "تمديد".

Segmentation or tokenization are commonly used interchangeably. The segmentation process is applied to segment the article into sentences and prepare for the next steps. We use the Natural Language Toolkit (NLTK) (Loper and Bird, 2002) to tokenize sentences and words. We are aware that some scholars weigh tokenization differently such as when tokenization breaks the words into constituent prefix(es), stem, and suffix(s) (Mubarak, 2017; Abdelali et al., 2016; El-Defrawy et al., 2015; Pasha et al., 2014). However, ATS lemmatization accomplishes the intended purpose of the other definition of Arabic tokenization.

Stop words have a major impact on text summarization because they impact the length of the articles and summaries, and increase the frequency of words which in both cases would change the weights of sentences (El-Khair, 2017; Al-Taani and Al-Omour, 2014). To remove the stop words, we used a list of stop words prepared by Abu El-khair et al (El-Khair, 2017) which contains 1,377 words.

For our evaluation, the final and most crucial step for preprocessing the text is lemmatization. This step can improve the accuracy of the summarization and evaluation process. Lemmatization is the process of reducing words to their basic root by removing the attached affixes of words. LANS dataset does not store the data in the lemmatized format, because lemmatization is usually used in the training or testing on the original data. Many lemmatizers are considered such as Alkhalil (Boudchiche and Mazroui, 2019), ISRI (Khoja) (El-Defrawy et al., 2015), Madamira (Pasha et al., 2014), CAMeL (Obeid et al., 2020), but only Farasa (Mubarak, 2017; Abdelali et al., 2016) is applied because it outperforms the state-of-the-art CAMel by a slight margin and its fast performance on large-scale datasets. Following all the mentioned steps, the dataset is passed for automatic evaluation (see sec 6).

4 LANS Description

LANS builds 8,443,484 articles and their summaries from 22 newspapers of 19 Arab countries dated from 1999 to 2019. The high-level overall statistics in Table 2 show that some newspapers have more data than the others. This does not undermine any country's newspapers. Among the

newspapers with a long history of journalism, most of them have been published on physical newspapers before newspapers become digitalized. The dates of collection reflect how much data is available in the e-archive for each newspaper. For instance, Gate Ahram newspaper from Egypt (Gat, 2022) is established in 1875 and has been published since then. However, the available e-archive for the newspaper starts from 2016. Each newspaper's webpage has its own e-archive and its own progress over time. This is why the variations of collection dates exist.

LANS encompasses 19 Arab countries for MSA diversity. One of the overlooked aspects of diversity in Arabic is the diversity of MSA in the Arab countries. It is true that all the newspapers in the Arab countries use the same MSA, but events, culture, and use of vocabulary are different from one country to another. Therefore, it is necessary to collect such diverse data from each country. To achieve MSA diversity in LANS, our dataset encompasses 19 Arab countries - except for the Comoros Islands, Djibouti, and Somalia because of the scarcity of data in their newspapers.

Further, LANS provides a wide-ranging topic variety. The collected data from each country covers different categories, and some newspapers have more categories than others, which enhances the diversity of categories in LANS. Some newspapers have only a few categories (not less than 7), while some others have more than 9 categories including local news, international, political, financial, society, sports, technology, art, health, and religious news articles. This category diversity is one of the features of LANS. It allows researchers to not only create subdatasets, but also create sub-subdataset of any of the subdatasets. For example, a subset can be all articles/summaries from Saudi Arabia. Then, a sub-subdataset can be the local news categories from the subset of Saudi Arabia articles/summaries. This type of diversity can be created from LANS.

The dataset is chunked into separate XML files, each file is under 2 GB to make it easier to load and process. The total size of the whole dataset is 32GB. Each country's dataset is a subset of the whole dataset, and researchers have the freedom to choose a subset or several subsets (by specific countries) to train and evaluate ATS models.

5 Experiment

Since the ATS field is still under-researched for *abstractive* summarization, it is difficult to achieve multiple comparisons among the available works. Therefore, we created a translate-summarize-translate pipeline from the available pretrained state-of-the-art multi-language models such mT5 (Xue et al., 2020), mBART (Tang et al., 2020), and CRISS (Tran et al., 2020). For our experiment, we chose mT5 becasue of its wide coverage of 101 languages and support for 41 languages. The model is utilized to generate summaries of the 1,000 randomly sampled articles, and then compare them with LANS ground-truth summaries using ROUGE-N. In a high-level description, the pipeline inputs the preprocessed samples as mentioned earlier in section 3.2.2, translates the articles (Arabic $\rightarrow$ English), generates summaries from the translated articles, then translates the generated summaries (English $\rightarrow$ Arabic) for evaluation. The model for each step of the pipeline will be given later.

Some of the pipeline steps to generate automatic text summaries are tuned to adapt Arabic language. Firstly, we preprocess the text, as detailed in section 3.2.2. Secondly, we translate the articles from Arabic to English. We apply OPUS-MT (Tiedemann and Thottingal, 2020) project. OPUS-MT is based on Marian-NMT (Junczys-Dowmunt et al., 2018), a state-of-the-art transformer-based Neural Machine Translation (NMT), and trained on OPUS data using OPUS-MT-Train. The translation achieves accurate results in machine translation. Next, since articles are translated into English, we process the articles to generate automatic text summaries using mT5 which inherits all the benefits of T5 (Raffel et al., 2019). The automatic text summaries currently are English. Finally, we translate automatic text summaries into Arabic by again applying the OPUS-MT project as described in the second step. An example of the ground-truth summary and a generated Arabic summary are displayed in Table 4.

Both summaries are evaluated by ROUGE (Ganesan, 2018) evaluation metric and will be used for human evaluation (see sec 6.2). We apply ROUGE-1, ROUGE-2, and ROUGE-L to consider different summary lengths. Moreover, we also show how lemmatization impacts the accuracy. The results are reported in Table 5. The results show that the summaries generated by mT5 achieve

Source	Summary
LANS	من المقرر الكشف عن اسماء اهم خمسة مرشحين لجائزة افضل لاعب كرة قدم في افريقيا للعام الحالي غداً الأحد ويتوقع ان يكون كابتن منتخب نيجيريا جاي جاي اوكوتشا من بين اقوى المرشحين للجائزة السنوية.
mT5-based pipeline	من المقرر الكشف عن أفضل خمسة مرشحين لأفضل لاعب كرة قدم في أفريقيا لهذا العام هذا الأحد، ومن المتوقع أن يكون الكابتن المنتخب لنيجيريا جاي أوكوكوشا من بين أقوى المرشحين للجائزة السنوية. أوكوشا، المرشح الرئيسي لأفضل لاعب أفريقي.

Table 4: Table presents a sample of two summaries from LANS and mT5-based pipeline.

lower scores before applying the lemmatization process. After we lemmatized the summaries by Farasa, the results improve by a good margin. In both cases, for a model that has not been designed for Arabic language, mT5 shows good scores when scored with LANS summaries see Table 4.

	Before Lemmatization			After Lemmatization		
	R-1	R-2	R-L	R-1	R-2	R-L
mT5	0.3	0.12	0.28	0.44	0.19	0.38

Table 5: Results of the generated summaries referenced to LANS summaries.

6 Intrinsic Evaluation of LANS

We apply two methods of evaluation to validate the reliability of the summaries from LANS. The first is an automatic evaluation which examines the summarization techniques in LANS. It uses the following metrics: *compression ratio*, *fragment density*, and *coverage*. The automatic evaluation has been performed on the whole dataset. The second evaluation is performed by experts which verifies the quality of LANS by randomly extracting 1,000 articles and their respective summaries, which are evaluated by experts.

6.1 Automatic Evaluation

To assess LANS, we apply 3 common metrics to quantify the abstractness of LANS's summaries and examine their strategies. Note that summaries can be *extractive* or *abstractive*; extractive summaries derive words from the source text, while abstractive summaries use novel words to describe the source text. The applied metrics used are *compression ratio, fragment density (abstractivity), and coverage* (Grusky et al., 2018; Bommasani and Cardie, 2020). **Compression Ratio** quantifies the conciseness of summaries, and is defined as the ratio of words between a summary and an article:

$$\mathbf{CMP}_w(S, A) = 1 - \frac{|S|}{|A|} \tag{1}$$

where $|S|$ is the summary's length and $|A|$ is the article's length in words. **Coverage** by (Grusky et al., 2018) quantifies how much the summary borrows words from the article. Its formula is below:

$$\mathbf{COV}(S, A) = \frac{1}{|S|} \sum_{t \in T(S,A)} |t| \tag{2}$$

where $T(S, A)$ is the set of extractive phrases in summary S extracted from article A, and t is the summary tokens (words) derived from the article. In abstractive summaries, it is preferred not to derive many words from the article.

Fragment Density is proposed by (Grusky et al., 2018), and later introduced as **Abstractivity** in (Bommasani and Cardie, 2020) with a slight change that generalizes it. This paper uses fragment density. It quantifies how well the summaries can construct a sequence of words that are greedily matched in the article. It is measured as the following:

$$\mathbf{DENS}(S, A) = \frac{1}{|S|} \sum_{t \in T(S,A)} |t|^2 \tag{3}$$

The results of the automatic evaluation are reported in Table 6. The ↓ arrow for coverage scores (COV) indicates how abstractive the summaries are from each source. The reported low scores signify that the summaries have novel words to describe the articles. The ↑ arrows for density (DENS) and fragment compression (CMP) mean the higher the better. The highest score for density is in Hespress(Morocco) newspaper summaries, and the lowest is in WAL (a Libyan news agency). For compression, the most concise summaries are reported from Alakhbar (Lebanon), and the least concise ones are reported from Alsudan Alyoom (Sudan). The diversity exists among the Arab countries' style of writing the summaries, and the indi-

Dataset	COV↓	DENS↑	CMP↑	Dataset	COV↓	DENS↑	CMP↑
Elkhabar(Algeria)	0.34	0.87	0.77	Alwatan(Oman)	0.35	0.64	0.68
Alwasat(Bahrain)	0.32	0.88	0.51	Alquds(Palestine)	0.28	0.74	0.65
Gate Ahram(Egypt)	0.27	0.81	0.57	Alquds-UK(Palestine)	0.39	0.90	0.79
Youm7(Egypt)	0.31	0.86	0.53	Alwatan(Qatar)	0.24	0.58	0.74
Aldustoor(Jordan)	0.25	0.52	0.50	Aljazira(Saudi Arabia)	0.23	0.46	0.57
Annahar(Kuwait)	0.24	0.57	0.72	Alryiadh(Saudi Arabia)	0.30	0.73	0.51
Almadapaper(Iraq)	0.45	0.52	0.64	Alsudan Alyoom(Sudan)	0.36	0.31	0.49
Alakhbar(Lebanon)	0.27	0.49	0.82	Zamanalwsl(Syria)	0.26	0.62	0.59
WAL(Libya)	0.32	0.30	0.55	Alssabah(Tunisia)	0.26	0.70	0.58
Sahara Media(Mauritania)	0.32	0.88	0.68	Albayan(Emirates)	0.41	0.35	0.65
Hespress(Morocco)	0.38	1.01	0.78	Almasdar(Yemen)	0.38	0.92	0.77

Table 6: Automatic evaluation results of LANS comparing all newspapers to each other. The up arrow ↑ indicates that higher is better and the opposite for the down arrow ↓. The results show the diversity among the collected datasets from one source to another. It also shows there is a high level of abstractiveness and conciseness.

cation of that is the varying scores in all metrics. The detailed distributions of *fragment density* and *coverage* across LANS dataset are displayed in the appendix Figure 2

6.2 Human Evaluation

Relying on only automatic evaluation and ROUGE metric may result in some limitations, such as biases in scoring against the systems that depend more on paraphrasing such as abstractive systems(Grusky et al., 2018). As a result, even though meaningful summaries are generated, ROUGE can be subjective and assigns a low score to well-generated summaries(See et al., 2017). Therefore, we conduct human evaluation.

Human evaluation is costly, but the results from the automatic method described in Sec. 6.1 are yet to be verified by experts. A survey is created for human experts to assess which summaries capture the full **key information of the articles**, have better **readability**, and have **syntactic correctness**. The survey contained the 1,000 random samples selected for the experiment in Sec. 5. Each survey question contains the following data: the full article; Choice 1: LANS summary; Choice 2: mT5-based generated summary; and Choice 3: none-of-the-above (non of the summaries). Choices 1 and 2 were shuffled and anonymized, so human experts can make fairer choices with less biases. For example, if Choice 1 was always LANS's summary, then human experts may form a judgement to always choose Choice 1. Therefore, the choices were shuffled. Besides, the choices were anonymous. It means that human evaluators do not know the origin of each summary.

The experts who did the survey are highly knowledgeable in Arabic. For a human expert to evaluate the survey; an expert should be an Arabic native speaker, also, an expert should at least have a bachelor's degree majoring in Arabic Language. The experts were asked not only to choose which choice is the fittest for the given criteria, but also to provide their feedback on the choices. Human evaluation results show that 954, out of the 1,000, LANS extracted summaries have more accurate semantic representation, and correct syntactic forms. The semantic representation means that the summary captures salient and key information of the article and has better readability. The results, also, show that 2 of the choices are "none", which means neither summaries meet the required criteria. While the ROUGE scores are low between the automatically generated summaries and the LANS summaries, the 95.4% approval rating for LANS summaries during the human evaluation validates the use of the descriptions present in the source code of the articles as their summaries.

7 Conclusion

This work presents LANS, a large-scale and diverse text summarization dataset of more than 8 million new articles paired with their summaries written by journalists. The summaries are collected from the metadata of 22 scraped popular Arab newspapers' websites from the period between 1999 to 2019. For each of those resources, LANS considered a wide range of topics. The work applied two evaluation methods (automatic and human) to verify the superiority of the extracted summaries in LANS. The dataset can be accessed upon request. [3]. LANS offers this dataset for researchers to advance the field of ATS, and takes advantage of the data to

[3]Request data from first author

train and evaluate the results of new models on this dataset.

8 Limitations

The distribution of data in LANS is far from uniform with regards to the newspapers coming from each country. This disparity is primarily driven by the varying number of newspapers in different countries. As a result, some nations' data representation is much more than others due to the former's extensive media landscape.

This uneven distribution underscores the importance of considering geographic and media-related factors when conducting data-driven research or analysis.

9 Ethical Statement

In accordance with ethical research practices, it is important to clarify that the data collection process for the LANS dataset did not violate any copyrights or intellectual property rights. The dataset comprises articles and their summaries obtained from publicly accessible websites of 22 major Arab newspapers, all of which span from 1999 to 2019. Every article included in the dataset is properly cited, including its originating source, and each has an associated URL, allowing for verification and direct reference. The data is solely utilized for academic and research purposes, intended to advance the field of Arabic Text Summarization (ATS). The extraction and use of this data adhere to all relevant ethical guidelines, ensuring that the journalistic integrity of the original articles and their authors is maintained. Thus, the dataset aims to serve as a high-quality and diverse resource for research while respecting all ethical and legal norms.

References

2022. Gate ahram newspaper (egypt). http://gate.ahram.org.eg/. Accessed: 2020-02-02.

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–16.

Abdullah M Abu Nada, Eman Alajrami, Ahmed A Al-Saqqa, and Samy S Abu-Naser. 2020. Arabic text summarization using arabert model using extractive text summarization approach. *International Journal of Academic Information Systems Research (IJAISR)*.

Fatima T AL-Khawaldeh and Venus W Samawi. 2015. Lexical cohesion and entailment based segmentation for arabic text summarization (lceas). *World of Computer Science & Information Technology Journal*, 5(3).

Molham Al-Maleh and Said Desouki. 2020. Arabic text summarization using deep learning approach. *Journal of Big Data*, 7(1):1–17.

Asma Bader Al-Saleh and Mohamed El Bachir Menai. 2016. Automatic arabic text summarization: a survey. *Artificial Intelligence Review*, 45(2):203–234.

Ahmad T Al-Taani and Maha M Al-Omour. 2014. An extractive graph-based arabic text summarization approach. In *The International Arab Conference on Information Technology*.

Abdullah Ayedh, Guanzheng Tan, Khaled Alwesabi, and Hamdi Rajeh. 2016. The effect of preprocessing on arabic document categorization. *Algorithms*, 9(2):27.

Riadh Belkebir and Ahmed Guessoum. 2015. A supervised approach to arabic text summarization using adaboost. In *New contributions in information systems and technologies*, pages 227–236. Springer.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096.

Mohamed Boudchiche and Azzeddine Mazroui. 2019. A hybrid approach for arabic lemmatization. *International Journal of Speech Technology*, 22(3):563–573.

Amina Chouigui, Oussama Ben Khiroun, and Bilel Elayeb. 2021. An arabic multi-source news corpus: Experimenting on single-document extractive summarization. *Arabian Journal for Science and Engineering*, 46(4):3925–3938.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2020. Gsum: A general framework for guided neural abstractive summarization. *arXiv preprint arXiv:2010.08014*.

Fouad Soufiane Douzidia and Guy Lapalme. 2004. Lakhas, an arabic summarization system. In *Proceedings of DUC*, volume 4, pages 128–135. Citeseer.

Mahmoud El-Defrawy, Yasser El-Sonbaty, and Nahla Belal. 2015. Enhancing root extractors using light stemmers. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation: Posters*, pages 157–166.

Fatma El-Ghannam and Tarek El-Shishtawy. 2014. Multi-topic multi-document summarizer. *arXiv preprint arXiv:1401.0640*.

Mahmoud El-Haj and Rim Koulali. 2013. Kalimat a multipurpose arabic corpus. In *Second Workshop on Arabic Corpus Linguistics (WACL-2)*, pages 22–25.

Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. 2010. Using mechanical turk to create a corpus of arabic summaries.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Ibrahim Abu El-Khair. 2017. Effects of stop words elimination for arabic information retrieval: a comparative study. *arXiv preprint arXiv:1702.01925*.

Asmaa Elsaid, Ammar Mohammed, Lamiaa Fattouh, and Mohamed Sakre. 2022. A comprehensive review of arabic text summarization. *IEEE Access*.

Hamzah Noori Fejer and Nazlia Omar. 2014. Automatic arabic text summarization using clustering and keyphrase extraction. In *Proceedings of the 6th International Conference on Information Technology and Multimedia*, pages 293–298. IEEE.

Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*.

Hamdy Mubarak. 2017. Build fast and accurate lemmatization for arabic. *arXiv preprint arXiv:1710.06700*.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the 12th language resources and evaluation conference*, pages 7022–7032.

Arfath Pasha, Mohamed Al-Badrashiny, Mona T Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *LREC*, volume 14, pages 1094–1101.

Dragomir Radev, Eduard Hovy, and Kathleen McKeown. 2002. Introduction to the special issue on summarization. *Computational linguistics*, 28(4):399–408.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

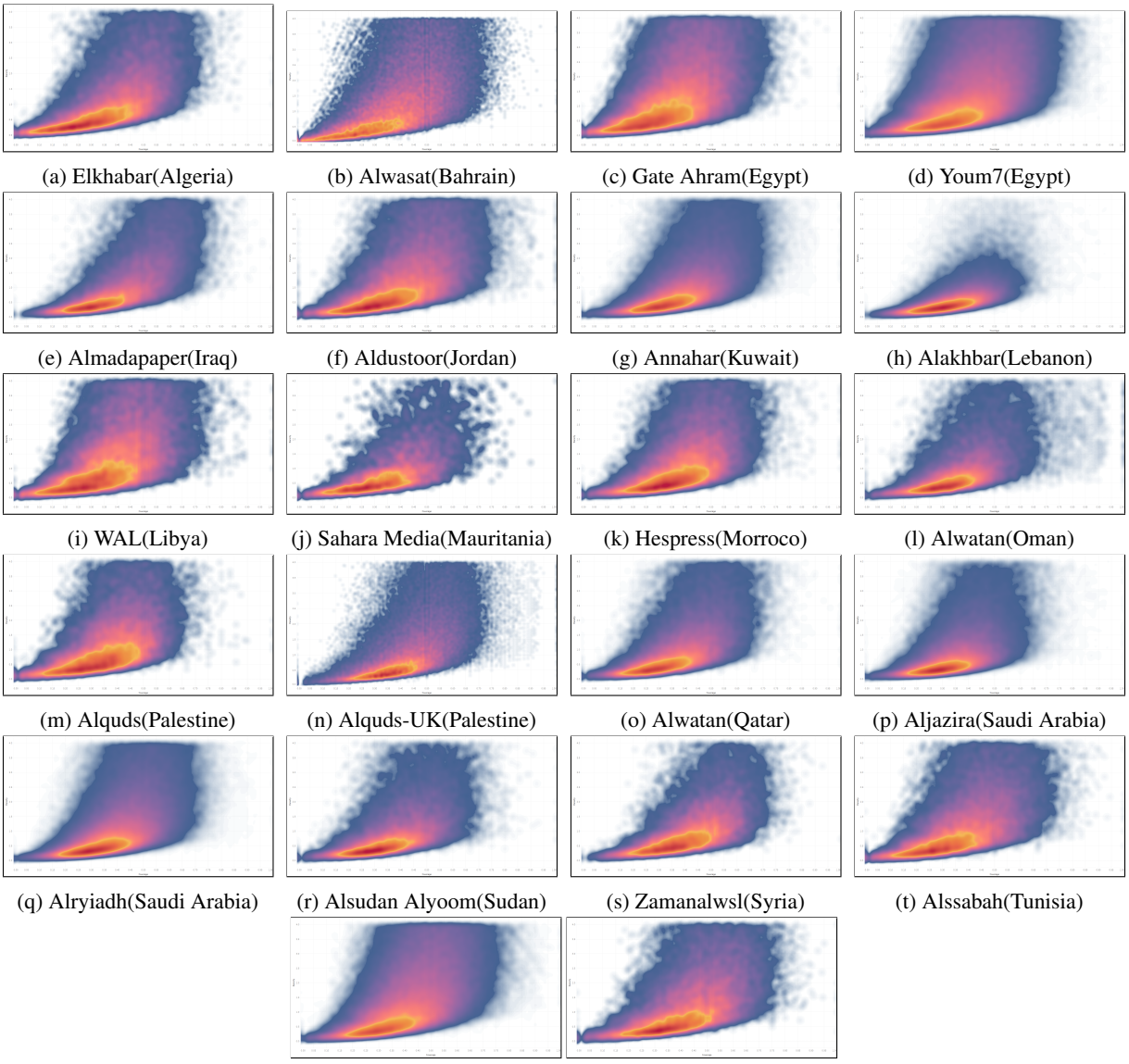
Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. *Advances in Neural Information Processing Systems*, 33:2207–2219.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

10 Appendix



(a) Elkhabar(Algeria) (b) Alwasat(Bahrain) (c) Gate Ahram(Egypt) (d) Youm7(Egypt)

(e) Almadapaper(Iraq) (f) Aldustoor(Jordan) (g) Annahar(Kuwait) (h) Alakhbar(Lebanon)

(i) WAL(Libya) (j) Sahara Media(Mauritania) (k) Hespress(Morroco) (l) Alwatan(Oman)

(m) Alquds(Palestine) (n) Alquds-UK(Palestine) (o) Alwatan(Qatar) (p) Aljazira(Saudi Arabia)

(q) Alryiadh(Saudi Arabia) (r) Alsudan Alyoom(Sudan) (s) Zamanalwsl(Syria) (t) Alssabah(Tunisia)

(u) Albayan(Emirates) (v) Almasdar(Yemen)

Figure 2: The distributions of fragment density and coverage across the datasets of LANS is displayed in the sub-figures. This shows how diverse the dataset is from one country to another. The sub-figures support table.6. Each sub-figure is a normalized bivariate density plot. The X-axis represents the coverage, and it ranges from 0 to 1. The Y-axis represents the Fragment density(Abstractiveness), and it ranges from 1 to 4. The red color shows where most of the summaries are, and the dark blue color indicates where the least summaries are. The extraction method is explained in section.6.1