MLSUM: The Multilingual Summarization Corpus

Thomas Scialom[*][‡]**, Paul-Alexis Dray**[*]**, Sylvain Lamprier**[‡]**,**
Benjamin Piwowarski[◊][‡]**, Jacopo Staiano**[*]
[◊] CNRS, France
[‡] Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
[*] reciTAL, Paris, France
`{thomas,paul-alexis,jacopo}@recital.ai`
`{sylvain.lamprier,benjamin.piwowarski}@lip6.fr`

Abstract

We present MLSUM, the first large-scale MultiLingual SUMmarization dataset. Obtained from online newspapers, it contains 1.5M+ article/summary pairs in five different languages – namely, French, German, Spanish, Russian, Turkish. Together with English news articles from the popular CNN/Daily mail dataset, the collected data form a large scale multilingual dataset which can enable new research directions for the text summarization community. We report cross-lingual comparative analyses based on state-of-the-art systems. These highlight existing biases which motivate the use of a multi-lingual dataset.

1 Introduction

The document summarization task requires several complex language abilities: understanding a long document, discriminating what is relevant, and writing a short synthesis. Over the last few years, advances in deep learning applied to NLP have contributed to the rising popularity of this task among the research community (See et al., 2017; Kryściński et al., 2018; Scialom et al., 2019). As with other NLP tasks, the great majority of available datasets for summarization are in English, and thus most research efforts focus on the English language. The lack of multilingual data is partially countered by the application of transfer learning techniques enabled by the availability of pre-trained multilingual language models. This approach has recently established itself as the *de-facto* paradigm in NLP (Guzmán et al., 2019).

Under this paradigm, for encoder/decoder tasks, a language model can first be pre-trained on a large corpus of texts in multiple languages. Then, the model is fine-tuned in one or more *pivot* languages for which the task-specific data are available. At inference, it can still be applied to the different languages seen during the pre-training. Because of the dominance of English for large scale corpora, English naturally established itself as a pivot for other languages. The availability of multilingual pre-trained models, such as BERT multilingual (M-BERT), allows to build models for target languages different from training data. However, previous works reported a significant performance gap between English and the target language, e.g. for classification (Conneau et al., 2018) and Question Answering (Lewis et al., 2019) tasks. A similar approach has been recently proposed for summarization (Chi et al., 2019) obtaining, again, a lower performance than for English.

For specific NLP tasks, recent research efforts have produced evaluation datasets in several target languages, allowing to evaluate the progress of the field in zero-shot scenarios. Nonetheless, those approaches are still bound to using training data in a *pivot* language for which a large amount of annotated data is available, usually English. This prevents investigating, for instance, whether a given model is as fitted for a specific language as for any other. Answers to such research questions represent valuable information to improve model performance for low-resource languages.

In this work, we aim to fill this gap for the automatic summarization task by proposing a large-scale MultiLingual SUMmarization (MLSUM) dataset. The dataset is built from online news outlets, and contains over 1.5M article-summary pairs in 5 languages: French, German, Spanish, Russian, and Turkish, which complement an already established summarization dataset in English.

The contributions of this paper can be summarized as follows:

1. We release the first large-scale multilingual summarization dataset;

2. We provide strong baselines from multilingual abstractive text generation models;

3. We report a comparative cross-lingual analysis of the results obtained by different approaches.

2 Related Work

2.1 Multilingual Text Summarization

Over the last two decades, several research works have focused on multilingual text summarization. Radev et al. (2002) developed MEAD, a multi-document summarizer that works for both English and Chinese. Litvak et al. (2010) proposed to improve multilingual summarization using a genetic algorithm. A community-driven initiative, MultiLing (Giannakopoulos et al., 2015), benchmarked summarization systems on multilingual data. While the MultiLing benchmark covers 40 languages, it provides relatively few examples (10k in the 2019 release). Most proposed approaches, so far, have used an extractive approach given the lack of a multilingual corpus to train abstractive models (Duan et al., 2019).

More recently, with the rapid progress in automatic translation and text generation, abstractive methods for multilingual summarization have been developed. Ouyang et al. (2019) proposed to learn summarization models for three low-resource languages (Somali, Swahili, and Tagalog), by using an automated translation of the New York Times dataset. Although this showed only slight improvements over a baseline which considers translated outputs of an English summarizer, results remain still far from human performance. Summarization models from translated data usually under-perform, as translation biases add to the difficulty of summarization.

Following the recent trend of using multi-lingual pre-trained models for NLP tasks, such as Multilingual BERT (M-BERT) (Pires et al., 2019)[1] or XLM (Lample and Conneau, 2019), Chi et al. (2019) proposed to fine-tune the models for summarization on English training data. The assumption is that the summarization skills learned from English data can transfer to other languages on which the model has been pre-trained. However a significant performance gap between English and the target language is observed following this process. This emphasizes the crucial need of multilingual training data for summarization.

2.2 Existing Multilingual Datasets

The research community has produced several multilingual datasets for tasks other than summarization. We report two recent efforts below, noting that both *i)* rely on human translations, and *ii)* only provide evaluation data.

The Cross-Lingual NLI Corpus The SNLI corpus (Bowman et al., 2015) is a large scale dataset for natural language inference (NLI). It is composed of a collection of 570k human-written English sentence pairs, associated with their label, entailment, contradiction, or neutral. The Multi-Genre Natural Language Inference (MultiNLI) corpus is an extension of SNLI, comparable in size, but including a more diverse range of text. Conneau et al. (2018) introduced the Cross-Lingual NLI Corpus (XNLI) to evaluate transfer learning from English to other languages: based on MultiNLI, a collection of 5,000 test and 2,500 dev pairs were translated by humans in 15 languages.

MLQA Given a paragraph and a question, the Question Answering (QA) task consists in providing the correct answer. Large scale datasets such as (Rajpurkar et al., 2016; Choi et al., 2018; Trischler et al., 2016) have driven fast progress.[2] However, these datasets are only in English. To assess how well models perform on other languages, Lewis et al. (2019) recently proposed MLQA, an evaluation dataset for cross-lingual extractive QA composed of 5K QA instances in 7 languages.

XTREME The Cross-lingual TRansfer Evaluation of Multilingual Encoders benchmark (Hu et al., 2020) covers 40 languages over 9 tasks. The summarization task is not included in the benchmark.

XGLUE To train and evaluate their performance across a diverse set of cross-lingual tasks, Liang et al. (2020) recently released XGLUE, covering both Natural Language Understanding and Generation scenarios. While no summarization task is included, it comprises a News Title Generation task: the data is crawled from a commercial news website and provided in form of article-title pairs for 5 languages (German, English, French, Spanish and Russian).

---

[1] https://github.com/google-research/bert/blob/master/multilingual.md

[2] For instance, see the SQuAD leaderboard: rajpurkar.github.io/SQuAD-explorer/

2.3 Existing Summarization datasets

We describe here the main available corpora for text summarization.

Document Understanding Conference Several small and high-quality summarization datasets in English (Harman and Over, 2004; Dang, 2006) have been produced in the context of the Document Understanding Conference (DUC).[3] They are built by associating newswire articles with corresponding human summaries. A distinctive feature of the DUC datasets is the availability of multiple reference summaries: this is a valuable characteristic since, as found by Rankel et al. (2013), the correlation between qualitative and automatic metrics, such as ROUGE (Lin, 2004), decreases significantly when only a single reference is given. However, due to the small number of training data available, DUC datasets are often used in a domain adaptation setup for models first trained on larger datasets such as Gigaword, CNN/DM (Nallapati et al., 2016; See et al., 2017) or with unsupervised methods (Dorr et al., 2003; Mihalcea and Tarau, 2004; Barrios et al., 2016a).

Gigaword Again using newswire as source data, the english Gigaword (Napoles et al., 2012; Rush et al., 2015; Chopra et al., 2016) corpus is characterized by its large size and the high diversity in terms of sources. Since the samples are not associated with human summaries, prior works on summarization have trained models to generate the headlines of an article, given its incipit, which induces various biases for learning models.

New York Times Corpus This large corpus for summarization consists of hundreds of thousand of articles from The New York Times(Sandhaus, 2008), spanning over 20 years. The articles are paired with summaries written by library scientists. Although (Grusky et al., 2018) found indications of bias towards extractive approaches, several research efforts have used this dataset for summarization (Hong and Nenkova, 2014; Durrett et al., 2016; Paulus et al., 2017).

CNN/Daily Mail One of the most commonly used dataset for summarization (Nallapati et al., 2016; See et al., 2017; Paulus et al., 2017; Dong et al., 2019), although originally built for Question Answering tasks (Hermann et al., 2015a). It consists of English articles from the CNN and The

Daily Mail associated with bullet point highlights from the article. When used for summarization, the bullet points are typically concatenated into a single summary.

NEWSROOM Composed of 1.3M articles (Grusky et al., 2018), and featuring high diversity in terms of publishers, the summaries associated with English news articles were extracted from the Web pages metadata: they were originally written to be used in search engines and social media.

BigPatent Sharma et al. (2019) collected 1.3 million U.S. patent documents, across several technological areas, using the Google Patents Public Datasets. The patents abstracts are used as target summaries.

LCSTS The Large Scale Chinese Short Text Summarization Dataset (Hu et al., 2015) is built from 2 million short texts from the Sina Weibo microblogging platform. They are paired with summaries given by the author of each text. The dataset includes 10k summaries which were manually scored by human for their relevance.

3 MLSUM

As described above, the vast majority of summarization datasets are in English. For Arabic, there exist the Essex Arabic Summaries Corpus (EASC) (El-Haj et al., 2010) and KALIMAT (El-Haj and Koulali, 2013); those comprise circa 1k and 20k samples, respectively. Pontes et al. (2018) proposed a corpus of few hundred samples for Spanish, Portuguese and French summaries. To our knowledge, the only large-scale non-English summarization dataset is the Chinese LCSTS (Hu et al., 2015). With the increasing interest for cross-lingual models, the NLP community have recently released multilingual evaluation datasets, targeting classification (XNLI) and QA (Lewis et al., 2019) tasks, as described in 2.2, though still no large-scale dataset is avaulable for document summarization.

To fill this gap we introduce MLSUM, the first large scale multilingual summarization corpus. Our corpus provides more than 1.5 millions articles in French (FR), German (DE), Spanish (ES), Turkish (TR), and Russian (RU). Being similarly built from news articles, and providing a similar amount of training samples per language (except for Russian), as the previously mentioned CNN/Daily Mail, it can effectively serve as a multilingual extension of the CNN/Daily Mail dataset.

---

[3] http://duc.nist.gov/

3.1 Collecting the Corpus

The CNN/Daily Mail (CNN/DM) dataset (see Section 2.3) is arguably the most used large-scale dataset for summarization. Following the same methodology, we consider news articles as the text input, and their paired *highlights/description* as the summary. For each language, we selected an online newspaper which met the following requirements: (1) Being a *generalist* newspaper: ensuring that a broad range of topics is represented for each language allows to minimize the risk of training topic-specific models, a fact which would hinder comparative cross-lingual analyses of the models; (2) Having a large number of articles in their public online archive; (3) Providing human written highlights/summaries for the articles that can be extracted from the HTML code of the web page.

After a careful preliminary exploration, we selected the online version of the following newspapers: Le Monde[4] (French), Süddeutsche Zeitung[5] (German), El Pais[6] (Spanish), Moskovskij Komsomolets[7] (Russian), and Internet Haber[8] (Turkish).

For each outlet, we gathered archived articles from 2010 to 2019. We applied one simple filter: all the articles shorter than 50 words or summaries shorter than 10 words are discarded, so as to avoid articles containing mostly audiovisual content. Each article was archived on the Wayback Machine,[9] allowing interested research to fully re-build (or extend) MLSUM. We distribute the dataset as a list of immutable snapshot URLs of the articles, along with the accompanying corpus-construction code,[10] allowing to replicate the parsing and preprocessing procedures we employed. This is due to legal reasons: the content of the articles is copyrighted and redistribution might be seen as infringing of publishing rights.[11]

We provide recommended train/validation/test splits following a chronological ordering based on the articles' publication dates. In our experiments below, we train/evaluate the models on the training/test splits obtained in this manner. Specifically,

---

[4] www.lemonde.fr
[5] www.sueddeutsche.de
[6] www.elpais.com
[7] www.mk.ru
[8] www.internethaber.com
[9] web.archive.org, using https://github.com/agude/wayback-machine-archiver
[10] https://github.com/recitalAI/MLSUM
[11] A similar approach has been adopted for several dataset releases in the recent past, such as Question Answering Corpus (Hermann et al., 2015b) or XSUM (Narayan et al., 2018a).

we use: data from 2010 to 2018, included, for training; data from 2019 (~10% of the dataset) for validation (up to May 2019) and test (May-December 2019). While this choice is arguably more challenging, due to the possible emergence of new topics over time, we consider it as the realistic scenario a successful summarization system should be able to deal with. Incidentally, this also brings the advantage of excluding most cases of leakage across languages: it prevents a model, for instance, from seeing a training sample describing an important event in one language, and then being submitted for inference a similar article in another language, published around the same time and dealing with the same event.

3.2 Dataset Statistics

We report statistics for each language in ML-SUM in Table 1, including those computed on the CNN/Daily Mail dataset (English) for quick comparison. MLSUM provides a comparable amount of data for all languages, with the exception of Russian with ten times less training samples. Important characteristics for summarization datasets are the length of articles and summaries, the vocabulary size, and a proxy for abstractiveness, namely the percentage of novel n-grams between the article and its human summary. From Table 1, we observe that Russian summaries are the shortest as well as the most abstractive. Coupled with the significantly lower amount of articles available from its online source, the task can be seen as more challenging for Russian than for the other languages in MLSUM.

4 Models

We experimented on MLSUM with the established models and baselines described below. Those include supervised and unsupervised methods, extractive and abstractive models. For all the experiments, we train models on a per-language basis. We used the recommended hyperparameters for all languages, in order to facilitate assessing the robustness of the models. We also tried to train one model with all the languages mixed together, but we did not see any significant difference of performance.

4.1 Extractive summarization models

Oracle Extracts the sentences, within the input text, that maximise a given metric (in our experiments, ROUGE-L) given the reference summary. It

	FR	DE	ES	RU	TR	EN
Dataset size	424,763	242,982	290,645	27,063	273,617	311,971
Training set size	392,876	220,887	266,367	25,556	249,277	287,096
Mean article length	632.39	570.6	800.50	959.4	309.18	790.24
Mean summary length	29.5	30.36	20.71	14.57	22.88	55.56
Compression Ratio	21.4	18.8	38.7	65.8	13.5	14.2
Novelty (1-gram)	15.21	14.96	15.34	30.74	28.90	9.45
Total Vocabulary Size	1,245,987	1,721,322	1,257,920	649,304	1,419,228	875,572
Occurring 10+ times	233,253	240,202	229,033	115,144	248,714	184,095

Table 1: Statistics for the different languages. *EN* refers to CNN/Daily Mail and is reported for comparison purposes. Article and summary lengths are computed in words. *Compression ratio* is computed as the ratio between article and summary length. *Novelty* is the percentage of words in the summary that were not in the paired article. *Total Vocabulary Size* is the total number of different words, and *Occurring 10+ times* is the total number of words occurring at least 10 times.

is an indication of the maximum one could achieve with extractive summarization. We rely on the implementation of Narayan et al. (2018b).

Random To compare the performances of the different models across languages, it is useful to include an unbiased model as a point of reference. To that purpose, we define a simple random extractive model that randomly extracts N words from the source document, with N fixed as the average length of the summary.

Lead-3 Simply selects the three first sentences from the input text. Sharma et al. (2019), among others, showed that this is a robust baseline for several summarization datasets such as CNN/DM, NYT and BIGPATENT.

TextRank An unsupervised algorithm proposed by Mihalcea and Tarau (2004). It consists in computing the co-similarities between all the sentences in the input text. Then, the most central to the document are extracted and considered as the summary. We used the implementation provided by Barrios et al. (2016b).

4.2 Abstractive summarization models

Pointer-Generator See et al. (2017) proposed the addition of the copy mechanism (Vinyals et al., 2015) on top of a sequence to sequence LSTM model. This mechanism allows to efficiently copy out-of-vocabulary tokens, leveraging attention (Bahdanau et al., 2014) over the input.

We used the publicly available OpenNMT implementation[12] with the default hyper-parameters.

However, to avoid biases, we limited the preprocessing as much as possible and did not use any sentence separators, as recommended for CNN/DM. This explains why we obtain lower ROUGE compared to when using a language-specific preprocessor.

M-BERT Encoder-decoder Transformer architectures are a very popular choice for text generation. Recent research efforts have adapted large pretrained self-attention based models for text generation (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019). In particular, Liu and Lapata (2019) added a randomly initialized decoder on top of BERT. Avoiding the use of a decoder, Dong et al. (2019) proposed to instead add a decoder-like mask during the pre-training to *unify* the language models for both encoding and generating. Both these approaches achieved SOTA results for summarization. In this paper, we only report results obtained following Dong et al. (2019), as in preliminary experiments we observed that a simple multilingual BERT (M-BERT), with no modification, obtained comparable performance on the summarization task.

5 Evaluation Metrics

ROUGE Arguably the most used set of metrics for Summarization, it is defined as the percentage of common *n*-grams between the evaluated summary and the human reference summary (Lin, 2004).

---

[12]opennmt.net/OpenNMT-py/Summarization.html

	FR	DE	ES	RU	TR	EN
Random	11.9 (7.5)	10.2 (6.7)	12.6 (6.5)	6.7 (2.5)	11.3 (6.3)	11.2 (10.6)
TextRank	12.6 (10.8)	13.3 (13)	9.5 (11.1)	3.3 (3.8)	21.5 (14.4)	28.6 (20.4)
Lead_3	19.7 (12.6)	33.1 (23.9)	13.7 (10.3)	5.9 (5.8)	28.9 (20.2)	35.2 (21.2)
Pointer Generator	23.6 (14.1)	35.1 (24.4)	17.7 (13.2)	5.7 (5.7)	32.6 (19.8)	33.3 (20.8)
M-BERT	25.1 (15.1)	42 (26.5)	20.4 (14.9)	9.5 (6.8)	32.9 (26.3)	35.4 (22.2)
Oracle	37.7 (24.7)	52.3 (31.7)	35.8 (26.5)	29.8 (20.3)	45.8 (26.4)	53.6 (30)

Table 2: ROUGE-L (METEOR) scores obtained by the models described in Section 4.1. *EN* refers to CNN/DM.

METEOR The Metric for Evaluation of Translation with Explicit ORdering (Banerjee and Lavie, 2005) was designed for the evaluation of machine translation output. It is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. METEOR is often reported in summarization papers (See et al., 2017; Dong et al., 2019) in addition to ROUGE.

Novelty Because of their use of copy mechanisms, some abstractive models have been reported to rely too much on extraction (See et al., 2017; Kryściński et al., 2018). Hence, it became a common practice to report the percentage of novel n-grams produced within the generated summaries.

Note that several approaches based on neural models have been recently proposed. Recent works (Eyal et al., 2019; Scialom et al., 2019) have proposed to evaluate summaries with QA based methods: the rationale is that a good summary should answer the most relevant questions about the article. Further, Kryściński et al. (2019) proposed a discriminator trained to measure the factualness of the summary. Böhm et al. (2019) learned a metric from human annotation. All these models were only trained on English datasets, preventing us to report them in this paper. The availability of ML-SUM will enable future works to build such metrics in a multilingual fashion.

6   Results and Discussion

The results we now present allow to compare the models across languages, and to investigate or hypothesize where their performance variations may come from.

To explain those, we note the following factors: *(1)* differences in the data, independently from the language, such as the structure of the article, the abstractiveness of the summaries, or the quantity; and *(2)* differences due to the language itself – either due to metric biases (e.g. sensitivity to morphology) or to biases inherent to the model. While the former have more to do with domain adaptation, the latter motivate further the development of multilingual datasets, since they are instrumental for studying such phenomena.

Turning to the observed results, we report in Table 2 the ROUGE-L and METEOR scores obtained by each model for all languages. We note that the overall order of systems (for each language) is preserved when using either metric (modulo some swaps between Lead_3 and Pointer Generator, but with relatively close scores).

Russian, the low-resource language in MLSUM For all experimental setups, the performance on Russian is comparatively low. This can be explained by at least two factors. First, the corpus is the most abstractive (see Table 1, limiting the performance obtained by extractive models (Random, LEAD-3, and Oracle). Second, one order of magnitude less training data is available for Russian than for the other MLSUM languages, a fact which can explain the impressive performance improvement (+66% in terms of ROUGE-L, see Table 2) between a *not pretrained* model (Pointer Generator) and a *pretrained* model (M-BERT).

How abstractive are the models? We report novelty in Figure 1. As previous works noted (See et al., 2017), pointer-generator networks are poorly abstractive, as they rely heavily on the copy mechanism. This is particularly true for Russian: the lack of data probably makes it easier to learn to copy than to cope with natural language generation. As expected, pretrained language models such as M-BERT are consistently and largely more abstractive, since they are exposed to other texts during pretraining.
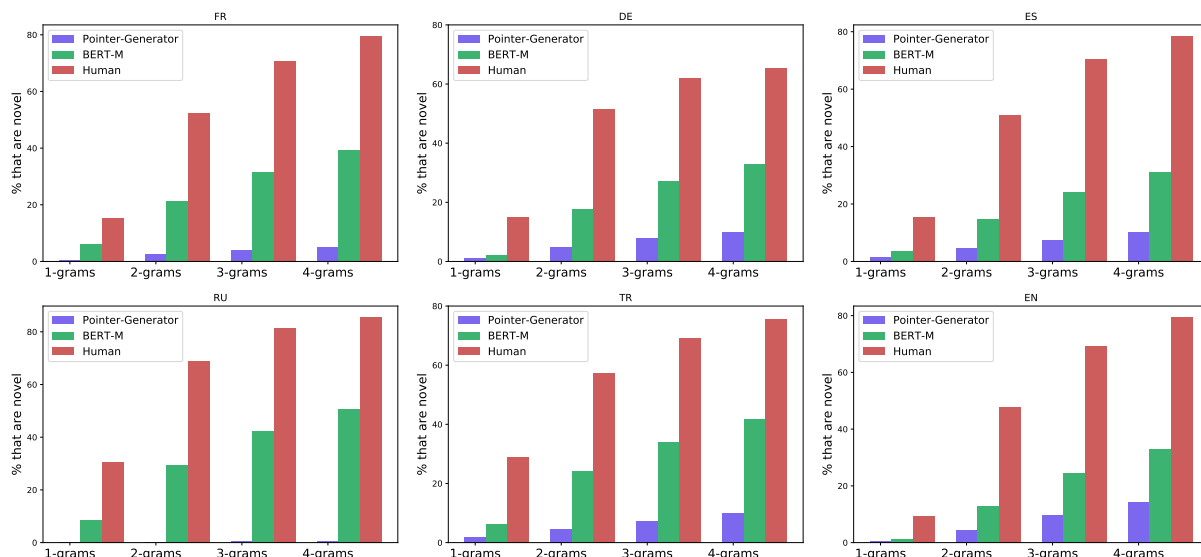
Figure 1: Percentage of novel n-grams for different abstractive models (neural and human), for the 6 datasets.

6.1 Model Biases toward Languages

Consistency among ROUGE scores The Random model obtains comparable ROUGE-L scores across all languages, except Russian. This can again be due to the Russian corpus characteristics – highest novelty, shortest summaries, and longest input documents (see Table 1). As this limits our ability to draw meaningful interpretations over Russian, compared to other languages, for the pair-wise language-based comparisons below we only focus on French, German, Spanish, and Turkish.

Abstractiveness of the datasets The Oracle performance can be considered as the upper limit for an extractive model since it extracts the sentences that provide the best ROUGE-L. We can observe that while being similar for English and German, and to some extent Turkish, the Oracle performance is lower for French or Spanish. However, as described in Figure 1, the percentage of novel words are similar for German (14.96), French (15.21) and Spanish (15.34). This may indicate that the relevant information to extract from the article is more spread among sentences for Spanish and French than for German. This is confirmed with the results of Lead-3: German and English have a much higher ROUGE-L – 35.20 and 33.09 – than French or Spanish – 19.69 and 13.70.

The case of TextRank The TextRank performance varies widely across the different languages, regardless Oracle. It is particularly surprising to see the low performance on German whereas, for this language, Lead-3 has a comparatively higher

	T/P	B/P
FR	0.53	1.06
DE	0.37	1.20
ES	0.53	1.15
RU	0.57	1.65
TR	0.65	1.01
CNN/DM (EN)	1.10	1.06
CNN/DM (EN, full preproc)	0.85	-
DUC (EN)	1.21	-
NEWSROOM (EN)	1.10	-

Table 3: Rouge-L ratios: T/P is the ratio of TextRank to Pointer-Generator and B/P is the ratio of M-BERT to Pointer-Generator. The results for CNN/DM-full pre-processing, DUC and NEWSROOM datasets are those reported in Table 2 of Grusky et al. (2018) (Pointer-C in their paper is our Pointer-Generator).

performance. On the other hand, the performance on English is remarkably high: the ROUGE-L is 33% higher than for Turkish, 126% higher than for French and 200% higher than for Spanish. We suspect that the TextRank parameters might actually overfit English. This is confirmed by the results reported in Table 3, where we report the performance ratio between TextRank and Pointer Generator on our corpus, as well as on CNN/DM and two other English corpora (DUC and NewsRoom). TextRank has a performance close to the Pointer Generator on English corpora (ratio between 0.85 to 1.21) but not in other languages (ratio between 0.37 to 0.65).
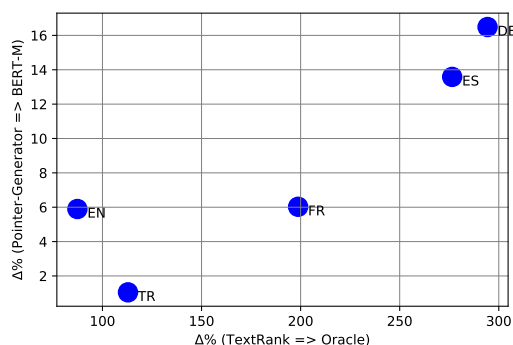
Figure 2: Improvement rates from TextRank to Oracle (in abscissa) against rates from Pointer Generator to M-BERT (in ordinate).

The benefits of pretraining We hypothesize that the closer an unsupervised model performance to its maximum limit, the less improvement would come from pretraining.

In Figure 2, we plot the improvement rate from TextRank to Oracle, against that of Pointer-Generator to M-BERT. Looking at the correlation emerging from the plot, the hypothesis appears to hold true for all languages, including Russian – not plotted for scaling reasons ($x = 808; y = 40$), with the exception of English. This is probably due to the aforementioned bias of TextRank towards the English language.

Pointer Generator and M-BERT We observe that M-BERT always outperforms the Pointer Generator. However, the ratio is not homogeneous across the different languages, see Table 3.

In particular, the improvement for German is much more important than the one for French. Interestingly, this observation is in line with the results reported for Machine Translation: the Transformer (Vaswani et al., 2017) outperforms significantly ConvS2S (Gehring et al., 2017) for English to German but obtains comparable results for English to French – see Table 2 in Vaswani et al. (2017). Neither model is pretrained, nor based on LSTM (Hochreiter and Schmidhuber, 1997), and they both use BPE tokenization (Shibata et al., 1999). Therefore, the main difference is represented by the self-attention mechanism introduced in the Transformer, while ConvS2S used only source-to-target attention.

We thus hypothesise that self-attention plays an important role for German but has a limited impact for French. This could find an explanation in the

morphology of the two languages: in statistical parsing, Tsarfaty et al. (2010) considered German to be very sensitive to word order, due to its rich morphology, as opposed to French. Among other reasons, the flexibility of its syntactic ordering is mentioned. This corroborates the hypothesis that self-attention might help preserving information for languages with higher degrees of word order freedom.

6.2 Possible derivative usages of MLSUM

Multilingual Question Answering Originally, CNN/DM was a Question Answering dataset (Hermann et al., 2015a), under the assumption that the information in the summary is also contained in the pair article. Questions can thus be generated from the summary sentences by masking the Named Entities contained therein. The masked entities represent the answers, and thus a masked question should be answerable given the source article.

As no multilingual QA *training* dataset has been proposed so far, this methodology could be applied to MLSUM to obtain a large-scale multilingual Question Answering corpus. Incidentally, this would also allow progressing towards multilingual Question Generation, a crucial component of the neural summarization metrics mentioned in Section 5.

News Title Generation While the release of MLSUM hereby described covers only article-summary pairs, the archived articles also include the titles. The accompanying code for parsing the articles allows to easily retrieve the titles and thus use them for News Title Generation.

Topic detection With the exception of Turkish, a topic/category can be associated with each article/summary pair, by simply parsing the corresponding URL. A natural application of this data for summarization would be for template based summarization (Perez-Beltrachini et al., 2019), using it as additional features. However, it can also be a useful multilingual resource for topic detection.

7 Conclusion

We presented MLSUM, the first large-scale Multi-Lingual SUMmarization dataset, comprising over 1.5M article/summary pairs in French, German, Russian, Spanish, and Turkish. We detailed its construction, and its complementary nature to the CNN/DM English summarization dataset.

We reported extensive preliminary experiments, highlighting biases observed in existing models, and analyzed the relative performances across languages of state-of-the-art models. In future works we plan to add other languages including Arabic and Hindi, and to investigate the adaptation of neural metrics to multilingual summarization.

Acknowledgements

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016a. Variations of the similarity function of textrank for automated summarization. *arXiv preprint arXiv:1602.03606*.

Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016b. Variations of the similarity function of textrank for automated summarization. *CoRR*, abs/1602.03606.

Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3101–3111, Hong Kong, China. Association for Computational Linguistics.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2019. Cross-lingual natural language generation via pre-training. *arXiv preprint arXiv:1909.10481*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Hoa Trang Dang. 2006. Duc 2005: Evaluation of question-focused summarization systems. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering*, pages 48–55. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*.

Bonnie Dorr, David Zajic, and Richard Schwartz. 2003. Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5*, pages 1–8. Association for Computational Linguistics.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3162–3172.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1998–2008.

M El-Haj, U Kruschwitz, and C Fox. 2010. Using mechanical turk to create a corpus of arabic summaries.

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

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. 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)*, pages 3938–3948.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org.

George Giannakopoulos, Jeff Kubina, John Conroy, Josef Steinberger, Benoit Favre, Mijail Kabadjov, Udo Kruschwitz, and Massimo Poesio. 2015. Multiling 2015: multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–274.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The flores evaluation datasets for low-resource machine translation: Nepali–english and sinhala–english. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6100–6113.

Donna Harman and Paul Over. 2004. The effects of human variation in duc summarization evaluation. In *Text Summarization Branches Out*, pages 10–17.

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

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015b. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems (NIPS)*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kai Hong and Ani Nenkova. 2014. Improving the estimation of word importance for news multi-document summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 712–721.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. Lcsts: A large scale chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1817.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark dataset for cross-lingual pretraining, understanding and generation. *arXiv preprint arXiv:2004.01401*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Marina Litvak, Mark Last, and Menahem Friedman. 2010. A new approach to improving multilingual summarization using a genetic algorithm. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 927–936. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3721–3731.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 95–100. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018a. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018b. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Jessica Ouyang, Boya Song, and Kathleen McKeown. 2019. A robust abstractive system for cross-lingual summarization. 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)*, pages 2025–2031.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Laura Perez-Beltrachini, Yang Liu, and Mirella Lapata. 2019. Generating summaries with topic templates and structured convolutional decoders. *arXiv preprint arXiv:1906.04687*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Elvys Linhares Pontes, Juan-Manuel Torres-Moreno, Stéphane Huet, and Andréa Carneiro Linhares. 2018. A new annotated portuguese/spanish corpus for the multi-sentence compression task. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Dragomir Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Arda Celebi, Hong Qi, Elliott Drabek, and Danyu Liu. 2002. Evaluation of text summarization in a cross-lingual information retrieval framework. *Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, Tech. Rep.*

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Peter A. Rankel, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2013. A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 131–136, Sofia, Bulgaria. Association for Computational Linguistics.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3237–3247, Hong Kong, China. Association for Computational Linguistics.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: A large-scale dataset for abstractive and coherent summarization. *arXiv preprint arXiv:1906.03741*.

Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. 1999. Byte pair encoding:

A text compression scheme that accelerates pattern matching. Technical report, Technical Report DOI-TR-161, Department of Informatics, Kyushu University.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kübler, Marie Candito, Jennifer Foster, Yannick Versley, Ines Rehbein, and Lamia Tounsi. 2010. Statistical parsing of morphologically rich languages (spmrl): what, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 1–12. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700.

A   Additional Metrics

	FR	DE	ES	RU	TR	EN
Random	17.95	13.08	17.34	7.68	15.92	17.55
TextRank	22.13	20.4	16.92	6.27	26.86	36.8
Lead_3	28.74	38.57	21.87	9.29	34.79	41.76
Point. Gen.	31.08	39.8	24.63	9.19	36.9	38.19
M-BERT	31.59	44.78	25.58	10.94	36.63	40.59
Oracle	47.32	57.23	45.23	36.14	50.61	59.62

Table 4: ROUGE-1 scores. *EN* refers to CNN/DM.

	FR	DE	ES	RU	TR	EN
Random	0.28	0.12	0.27	0.07	0.29	0.27
TextRank	6.65	6.94	4.85	1.3	12.59	13.61
Lead_3	9.84	25.66	6.25	1.54	20.0	16.65
Point. Gen.	10.12	25.96	6.54	1.18	21.77	14.38
M-BERT	10.61	30.75	8.61	1.75	20.15	16.39
Oracle	25.95	39.72	26.21	19.88	33.55	32.87

Table 5: ROUGE-2 scores. *EN* refers to CNN/DM.

B   Topic Shift

With the exception of Turkish, the article URLs in MLSUM allow to identify a category for a given article. In Figure 3 we show the shift over categories among time. In particular, we plot the 6 most frequent categories per language.
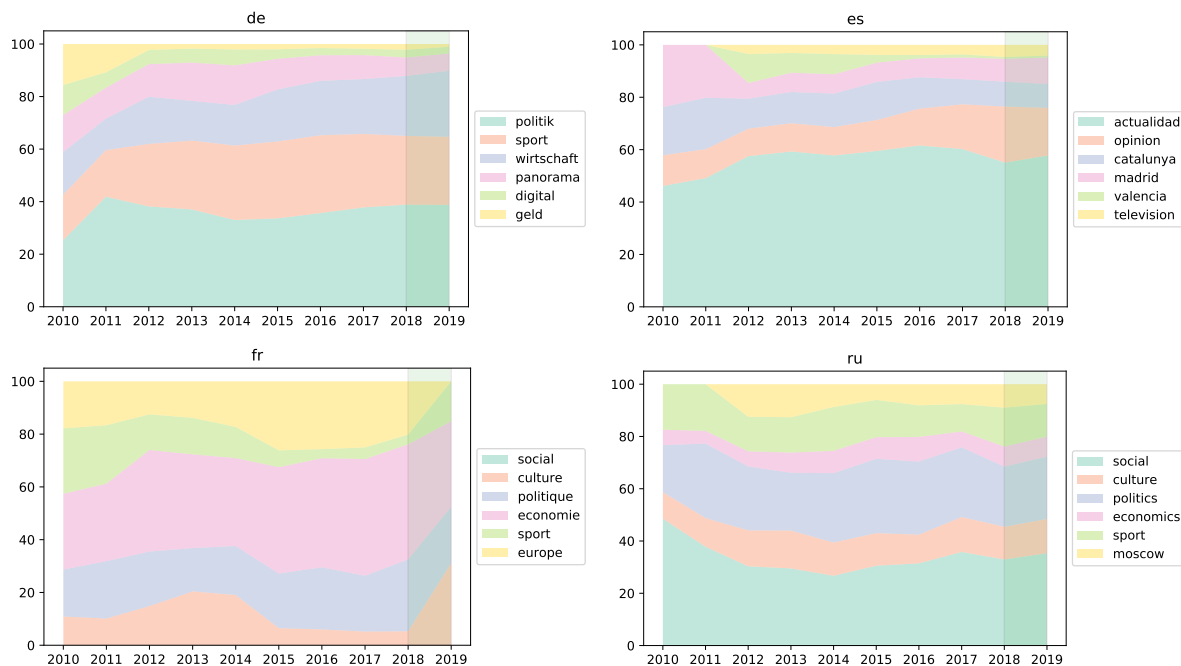

Figure 3: Distribution of topics for German (top-left), Spanish (top-right), French (bottom-left) and Russian (bottom-right), grouped per year. The shaded area for 2019 highlights validation and test data.

C Samples

– FRENCH –

summary Terre d'origine du clan Karzaï, la ville méridionale de Kandahar est aussi un bastion historique des talibans, où le mollah Omar a vécu et conservé de profondes racines. C'est sur cette terre pachtoune, plus qu'à Kaboul, que l'avenir à long terme du pays pourrait se décider.

body Lorsque l'on parle de l'Afghanistan, les yeux du monde sont rivés sur sa capitale, Kaboul. C'est là que se concentrent les lieux de pouvoir et où se détermine, en principe, son avenir. C'est aussi là que sont réunis les commandements des forces civiles et militaires internationales envoyées sur le sol afghan pour lutter contre l'insurrection et aider le pays à se reconstruire. Mais, à y regarder de plus près, Kaboul n'est qu'une façade. Face à un Etat inexistant, une structure du pouvoir afghan encore clanique, des tribus restées puissantes face à une démocratie artificielle importée de l'extérieur, la vraie légitimité ne vient pas de Kaboul. La géographie du pouvoir afghan aujourd'hui oblige à dire qu'une bonne partie des clés du destin de la population afghane se trouve au sud, en terre pachtoune, dans une cité hostile aux étrangers, foyer historique des talibans, Kandahar. Kandahar est la terre d'origine du clan Karzaï et de sa tribu, les Popalzaï. Hamid Karzaï, président afghan, tient son pouvoir du poids de son clan dans la région. Mi-novembre 2009, dans la grande maison de son frère, Wali, à Kandahar, se pressaient des chefs de tribu venus de tout l'Afghanistan, les piliers de son réseau. L'objet de la rencontre : faire le bilan post-électoral après la réélection contestée de son frère à la tête du pays. Parfois décrié pour ses liens supposés avec la CIA et des trafiquants de drogue, Wali Karzaï joue un rôle politique méconnu. Il a organisé la campagne de son frère, et ce jour-là, à Kandahar, se jouait, sous sa houlette, l'avenir de ceux qui avaient soutenu ou au contraire refusé leur soutien à Hamid. Chef d'orchestre chargé du clan du président, Wali est la personnalité forte du sud du pays. Les Karzaï adossent leur influence à celle de Kandahar dans l'histoire de l'Afghanistan. Lorsque Ahmad Shah, le fondateur du pays, en 1747, conquit la ville, il en fit sa capitale. "Jusqu'en 1979, lors de l'invasion soviétique, Kandahar a incarné le mythe de la création de l'Etat afghan, les Kandaharis considèrent qu'ils ont un droit divin à diriger le pays", résume Mariam Abou Zahab, experte du monde pachtoune. "Kandahar, c'est l'Afghanistan, explique à ceux qui l'interrogent Tooryalaï Wesa, gouverneur de la province. La politique s'y fait et, encore aujourd'hui, la politique sera dictée par les événements qui s'y dérouleront." Cette emprise de Kandahar s'évalue aux places prises au sein du gouvernement par "ceux du Sud". La composition du nouveau gouvernement, le 19 décembre, n'a pas changé la donne. D'autant moins que les rivaux des Karzaï, dans le Sud ou ailleurs, n'ont pas réussi à se renforcer au cours du dernier mandat du président. L'autre terre pachtoune, le grand Paktia, dans le sud-est du pays, à la frontière avec le Pakistan, qui a fourni tant de rois, ne dispose plus de ses relais dans la capitale. Kandahar pèse aussi sur l'avenir du pays, car s'y trouve le coeur de l'insurrection qui menace le pouvoir en place. L'OTAN, défiée depuis huit ans, n'a cessé de perdre du terrain dans le Sud, où les insurgés contrôlent des zones entières. Les provinces du Helmand et de Kandahar sont les zones les plus meurtrières pour la coalition et l'OTAN semble dépourvue de stratégie cohérente. Kandahar est la terre natale des talibans. Ils sont nés dans les campagnes du Helmand et de Kandahar, et le mouvement taliban s'est constitué dans la ville de Kandahar, où vivait leur chef spirituel, le mollah Omar, et où il a conservé de profondes racines. La pression sur la vie quotidienne des Afghans est croissante. Les talibans suppléent même le gouvernement dans des domaines tels que la justice quotidienne. Ceux qui collaborent avec les étrangers sont stigmatisés, menacés, voire tués. En guise de premier avertissement, les talibans collent, la nuit, des lettres sur les portes des "collabos". "La progression talibane est un fait dans le Sud, relate Alex Strick van Linschoten, unique spécialiste occidental de la région et du mouvement taliban à vivre à Kandahar sans protection. L'insécurité, l'absence de travail poussent vers Kaboul ceux qui ont un peu d'éducation et de compétence, seuls restent les pauvres et ceux qui veulent faire de l'argent." En réaction à cette détérioration, les Américains ont décidé, sans l'assumer ouvertement, de reprendre le contrôle de situations confiées officiellement par l'OTAN aux Britanniques dans le Helmand et aux Canadiens dans la province de Kandahar. Le mouvement a été progressif, mais, depuis un an, les Etats-Unis n'ont cessé d'envoyer des renforts américains, au point d'exercer aujourd'hui de fait la direction des opérations dans cette région. Une tendance qui se renforcera encore avec l'arrivée des troupes supplémentaires promises par Barack Obama. L'histoire a montré que, pour gagner en Afghanistan, il fallait tenir les campagnes de Kandahar. Les Britanniques l'ont expérimenté de façon cuisante lors de la seconde guerre anglo-afghane à la fin du XIXe siècle et les Soviétiques n'en sont jamais venus à bout. "On sait comment cela s'est terminé pour eux, on va essayer d'éviter de faire les mêmes erreurs", observait, mi-novembre, optimiste, un officier supérieur américain.

summary Die Wurzeln des Elends liegen in der Vergangenheit. Haiti bezahlt immer noch für seine Befreiung vor 200 Jahren. Auch damals nahmen die Wichtigen der Welt den Insel-Staat nicht ernst.

body Das Portrait von 1791 zeigt Haitis Nationalhelden François-Dominique Toussaint L'Ouverture. Er war einer der Anführer der Revolution in Haiti und Autor der ersten Verfassung. Die Wurzeln des Elends liegen in der Vergangenheit. Haiti bezahlt immer noch für seine Befreiung vor 200 Jahren. Auch damals nahmen die Wichtigen der Welt den Insel-Staat nicht ernst. Am vergangenen Wochenende schickte der britische Architekt und Gründer der Organisation Architecture for Humanity eine atemlose, verzweifelte E-Mail an seine Freunde und Unterstützer. "Nicht Erdbeben, sondern Gebäude töten Menschen" schrieb er in die Betreffzeile. Damit brachte er auf den Punkt, was auch der Geologe und Autor Simon Winchester oder der Urbanist Mike Davis immer wieder geschrieben haben - es gibt keine Naturkatastrophen. Es gibt nur gewaltige Naturereignisse, die tödliche Folgen haben. Die Konsequenz aus dieser Schlussfolgerung ist die Schuldfrage. Einfach lässt sie sich beantworten: Gier und Korruption sind fast immer die Auslöser einer Katastrophe. In Haiti aber liegen die Wurzeln der Tragödie tief in der Geschichte des Landes. Diese begann nach europäischer Rechnung im Jahre 1492, als Christopher Kolumbus auf der Insel landete, die ihre Ureinwohner Aytí nannten. Kolumbus benannte die Insel in Hispaniola um und gründete mit den Trümmern der gestrandeten Santa Maria die erste spanische Kolonie in der Neuen Welt. Ende des 17. Jahrhunderts besetzten französische Siedler den Westen der Insel, den Frankreich 1691 zur französischen Kolonie Sainte Domingue erklärte. Ideale der Französischen Revolution Gut hundert Jahre währte die Herrschaft der beiden Kolonialherren über die geteilte Insel. "Saint Domingue war die reichste europäische Kolonie in den Amerikas", schrieb der Historiker Hans Schmidt. 1789 kam fast die Hälfte des weltweit produzierten Zuckers aus der französischen Kolonie, die auch in der Produktion von Kaffee, Baumwolle und Indigo Weltmarktführer war. 450000 Sklaven arbeiteten auf den Plantagen, und sie erfuhren bald vom neuen Geist ihrer Herren. Die Französische Revolution brachte die Ideale von Freiheit, Gleichheit und Brüderlichkeit in die Karibik. Im August 1791 war es so weit. Der Voodoo-Priester Dutty Boukman rief während einer Messe zum Aufstand. Einer der erfolgreichsten Kommandeure der Rebellion war der ehemalige Sklave François-Dominique Toussaint L'Ouverture, nach dem heute der Flughafen von Port-au-Prince benannt ist. 1801 gab Toussaint dem Land seine erste Verfassung, die gleichzeitig eine Unabhängigkeitserklärung war. Für Napoleon sollte Haiti eine Schmach bleiben. Daraufhin sandte Napoleon Bonaparte Kriegsschiffe und Soldaten. Toussaint wurde verhaftet und nach Frankreich gebracht, wo er im Kerker starb. Doch als Napoleon im Jahr darauf die Sklaverei wieder einführen wollte, kam es erneut zum Aufstand. Verzweifelt baten die französischen Truppen im Sommer 1803 um Verstärkung. Da aber hatte Napoleon schon das Interesse an der Neuen Welt verloren. Im April hatte er seine Kolonie Louisiana an die Nordamerikaner verkauft, ein Gebiet, das rund ein Viertel des Staatsgebietes der heutigen USA umfasste. Für Napoleon sollte Haiti eine Schmach bleiben. Am 1. Januar 1804 erklärte der Rebellenführer Jean-Jacques Dessalines, die ehemalige Kolonie heiße nun Haiti und sei eine freie Republik. Der erste und bis zur Abschaffung der Sklaverei einzige erfolgreiche Sklavenaufstand der Neuen Welt war ein Schock für die Großmächte der Kolonialära, die ihren Reichtum auf der Sklaverei gegründet hatten. Ein Handel, der die Geschichte Haitis bis heute bestimmt Die Freiheit hatte ihren Preis. Ein Großteil der Plantagen war zerstört, ein Drittel der Bevölkerung Haitis den Kämpfen zum Opfer gefallen. Vor allem aber wollte keine Kolonialmacht die junge Republik anerkennen. Im Gegenteil -die meisten Länder unterstützten das Embargo der Insel und die Forderungen französischer Sklavenherren nach Reparationszahlungen. In der Hoffnung, als freie Nation Zugang zu den Weltmärkten zu erhalten, ließ sich die neue Machtelite Haitis auf einen Handel ein, der die Geschichte der Insel bis heute bestimmt. Mehr als zwei Jahrzehnte nach dem Sieg der Rebellen entsandte König Karl X. seine Kriegsschiffe nach Haiti. Ein Emissär stellte die Regierung vor die Wahl: Haiti sollte für die Anerkennung als Staat 150 Millionen Francs bezahlen. Sonst würde man einmarschieren und die Bevölkerung erneut versklaven. Haiti nahm Schulden auf und bezahlte. Bis zum Jahre 1947 lähmte die Schuldenlast die haitianische Wirtschaft und legte den Grundstein für Armut und Korruption. 2004 ließ der damalige haitianische Präsident Jean-Bertrand Aristide errechnen, was diese "Reparationszahlungen" für Haiti bedeuteten. Rund 22 Milliarden amerikanische Dollar Rückzahlung forderten seine Anwälte damals von der französischen Regierung. Vergebens. Lesen Sie auf der nächsten Seite, wie Haiti von den Akteuren der Weltbühne geschnitten wurde.

– SPANISH –

summary El aeropuerto ha estado hasta las 15.00 con sólo dos pistas por ausencia de 5 de los 18 controladores aéreos.-Varias aerolíneas han denunciado demoras de "hasta 60 minutos con los pasajeros embarcados"

body El espacio hará un repaso cronológico de la vida de la Esteban desde el momento en el que una completa desconocida comenzó a aparecer en los medios en 1998 como la novia de Jesulín de Ubrique hasta llegar a hoy en día, convertida en la princesa del pueblo, en concreto del popular madrileño distrito de San Blas donde vive, tal y como algunos la han calificado, y protagonista de portadas de revistas, diarios y portales web y de aparecer incluso entre los personajes más populares de Google. Junto a María Teresa Campos, estarán en el plató Patricia Pérez, presentadora del programa matinal de los sábados en Telecinco Vuélveme loca, quien ha conducido las campanadas en cuatro ocasiones, y los comentaristas Maribel Escalona, Emilio Pineda y José Manuel Parada.Los vuelos han venido registrando este viernes importantes retrasos en Barajas a pesar de que desde las 15.00 el aeropuerto opera con las cuatro pistas, según han informado fuentes de AENA, mientras las compañías han denunciado demoras por parte de los controladores de hasta 60 minutos con los pasajeros embarcados. Según los datos facilitados por AENA, la ausencia por la mañana de 5 de los 18 controladores que estaban programados en el turno de la torre de control de Barajas obligó a cerrar dos de las pistas del aeropuerto, lo que generó retrasos medios de 30 minutos.

– TURKISH –

summary Ataması yapılmayan öğretmenler miting yaptı. Öğretmen adaylarına Muharrem İnce ve TEKEL işçileri de destek verdi.

body Yetersiz açılan kadrolar nedeniyle ataması yapılamayan öğretmen adayları Ankara'da miting yaptı. Tekel işçilerinin de destek verdiği öğretmen adaylarının mitinginde öğretmen kökenli CHP Milletvekili Muharrem İnce de hazır bulundu. Türkiye'nin çeşitli illerinden gelen "Ataması Yapılmayan Öğretmenler Platformu" üyesi sözleşmeli öğretmenler, öğle saatlerinde Abdi İpekçi Parkı'nda toplandı. "Milletvekilliği için KPSS getirilsin", "1 kadrolu öğretmen = 3 ücretli öğretmen" ve "Ücretli köle olmayacağız" yazılı dövizler taşıyan ve aynı içerikli sloganlar atan öğretmenlerin düzenlediği mitinge, bazı siyasi parti, sivil toplum kuruluşu temsilcileri ve TEKEL işçileri de destek verdi. CHP Yalova Milletvekili Muharrem İnce, okullarda derslerin boş geçtiğini öne sürerek, "Okullar öğretmensiz, öğretmenler ise işsiz" dedi. Hükümetin bu gençlerin sesini duyması gerektiğini belirten İnce, "Bu ülkenin 250 bin eğitim fakültesi mezunu genci iş bekliyorsa bu hükümetin ve ülkenin ayıbıdır. Eğitim sorununu çözememiş bir hükümet bu ülkenin hiçbir sorununu çözememiş demektir. Bu kadar önemli bir soruna kulaklarını tıkayamaz" diye konuştu. "Ankara'nın göbeğinde derslerin boş geçtiğini" ileri süren İnce, "Bu ülkede fizik ve matematik öğretmeni atanmıyor ama bunların 100 katı din dersi öğretmeni atanıyor" dedi. Platform adına yapılan açıklamada da Türkiye'de her yıl üniversite bitirerek diplomasını alan öğretmenlerin eğitim alanındaki yetersizlik dolayısıyla işsizler kervanına katıldığı ifade edildi. Talep edilen hakların insancıl ve makul olduğu belirtilen açıklamada, öğretmenlerin haklarını vermeyenlerin kötü niyetli olduğu öne sürüldü. Açıklamada, hükümetin eğitim politikası eleştirilerek, sözleşmeli öğretmenlerin kadrolu atamalarının yapılması, öğretmen yetiştiren fakültelere öğretmen ihtiyacı kadar öğretmen adayı alınması ve KPSS yerine daha şeffaf bir atama sistemi getirilmesi istendi. ÖLÜM ORUCU BAŞLATACAKLAR Çeşitli sivil toplum kuruluşu temsilcilerinin de konuştuğu mitingde, kadrolu atamalar yapılmadığı takdirde iş bırakma eylemi ve ölüm orucu yapılacağı duyuruldu.

– RUSSIAN –

summary Старший преподаватель института коммунального хозяйства и строительства был задержан на днях в Москве за растление школьника

body Сладострастник в течение трех лет преследовал подростка в надежде совратить его. Как сообщили "МК" в следственном отделе по Хорошевскому району СУ СК при Прокуратуре РФ по Москве, 26 августа 2006 года 13-летний Павел вместе с другом отдыхал на берегу Москвы–реки рядом с Крылатским мостом. Там к ребятам подошел мужчина. Новый знакомый представился Евгением и предложил вместе пообедать в ресторане быстрого питания, а потом искупаться. Именно там, на берегу, педагог начал приставать к мальчику. Школьник убежал, но педофил успел снять голого подростка на мобильный телефон. После этого жизнь мальчика превратилась в сущий ад. Евгений узнал, где живет Павел, и стал шантажировать его. Этот кошмар продолжался три года. Преподаватель угрожал показать фотографию друзьям и знакомым Павла. Негодяй исписал непотребными надписями стены подъезда, где проживали друзья школьника. В один из дней он приехал в Сергиев Посад, к бабушке мальчика, и там накинулся на школьника с ножом. Наконец, отчаявшийся подросток рассказал обо всем матери, и та обратилась в милицию. Первый раз стражи порядка упустили 43-летнего педофила (к слову, он разведен). Милиционеры нагрянули в его квартиру, а Евгений под предлогом прощания с мамой-инвалидом зашел в соседнюю комнату, выпрыгнул из окна и был таков. Задержать извращенца стражам порядка помогла случайность. Несколько дней назад мать Павла увидела Евгения на станции метро "Александровский сад". На мужчине красовался парик, но женщина узнала негодяя в толпе. Милиционеры, дежурившие на станции, по просьбе дамы задержали растлителя. В его кармане они обнаружили нож. Кстати, за несколько дней до задержания мужчина подкараулил Павла на улице и ударил его по лицу, обзывая нехорошими словами. Всех, чьи дети стали жертвами педофила, просят звонить по телефону 8-499-197-88-17 или 02.