The State and Fate of Summarization Datasets

Anonymous ACL submission

Abstract

Automatic summarization has consistently attracted attention, due to its versatility and wide application in various downstream tasks. Despite its popularity, we find that annotation efforts have largely been disjointed, and have lacked common terminology. Consequently, it is challenging to discover existing resources or identify coherent research directions. To address this, we survey a large body of work spanning 133 datasets in over 100 languages, creating a novel ontology covering sample properties, collection methods and distribution. With this ontology we make key observations, including the lack in accessible high-quality datasets for low-resource languages, and the field's over-reliance on the news domain and on automatically collected distant supervision. Finally, we make available a web interface that allows users to interact and explore our ontology and dataset collection, as well as a template for *summarization data card*, which can be used to streamline future research into a more coherent body of work.[1]

1 Introduction

Summarization is the task of shortening a text while preserving the most important information it contains. This task has been explored for over 60 years (Luhn, 1958) yet it remains a longstanding challenge, as evidenced by the steadily increasing number of summarization datasets released every year (see Figure 1).

Despite the many research efforts in creating summarization datasets, we find they lack standardization and common terminology. This shortcoming makes it difficult to discover existing resources and bottlenecks, and to identify general directions for future research.

In this work, we conduct the most comprehensive survey of summarization datasets, covering a wide collection of 133 datasets across 104 languages. In Section 3, we develop a comprehensive ontology geared towards modern summarization datasets, positioning each dataset along seven different scales, including information about the samples (language, language modality, domain and summarization shape) as well as collection methods (annotation efforts and sources of supervision), and the manner in which the dataset is distributed.

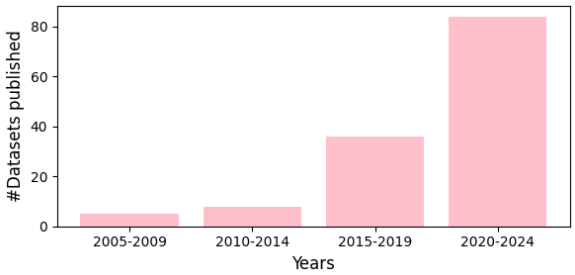

Figure 1: There is a continuous increase in the number of summarization datasets being published. However, we find that the field is missing standardization and common terminology.

In Section 4 we show that our proposed ontology surfaces five major trends and gaps in the field. First, we find that the lack of common terminology thus far often leads to inconsistency and to the omission of important details when presenting datasets, making comparison challenging. An important example of this inconsistency is the traditional abstractive-extractive distinction, which we argue have gradually shifted away from a binary trait to a more nuanced spectrum.

Second, we observe that the field is over-reliant on the news domain, which acts as a double-edged sword – being both easy to obtain and prone to low quality. This is especially evident for low-resource languages, where the vast majority of datasets are based on news articles.

Third, we find that the datasets available for low-resource languages tend to be of subpar standard

[1]We will make our resources publicly available.

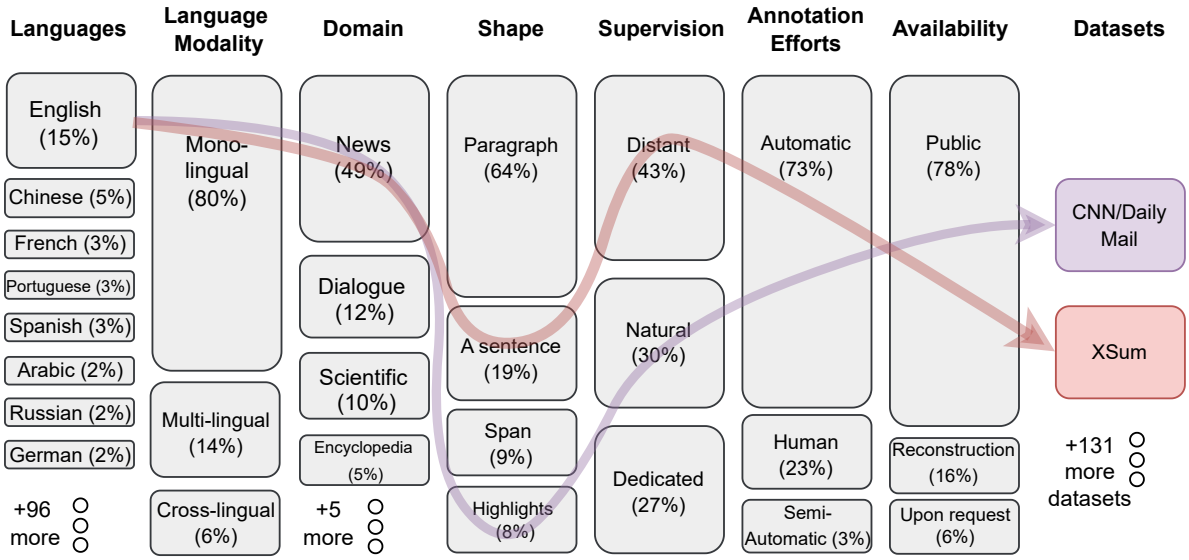

Figure 2: Our ontology for summarization datasets, accompanied by our annotations. The percentages in the language column indicate the proportion of datasets that support each language, where we count each multilingual dataset as multiple monolingual datasets. The arrows showcase the decision pathways for selecting specific datasets.

and especially challenging to discover. This is because many languages are currently supported only through multilingual datasets, which tend to compromise quality for diversity (Urlana et al., 2022), while also making it harder to know which languages are annotated in which dataset.

Fourth, we find that most of the annotations were not specifically created to serve as summaries.

Fifth, we observe that copyright concerns, especially in the news domain have been changing data distribution, with many datasets move from publishing the data towards publishing scripts which reconstruct it from URLs. We argue that this practice is problematic, as URLs may change over time. Furthermore, a few seminal summarization datasets that were published publicly have been either taken down or are considered legally questionable due to copyright violations.

In Section 5 we publish two resources to standardize and help organize the vast research in the field. First, we make our dataset collection publicly available through a web UI which allows users to search datasets matching different criteria according to our terminology, for example, retrieving all manually created datasets in a certain language. This UI serves a central location where researchers can discover datasets, as well as identify new research directions. We welcome community contributions to keep this resource updated with newly published research.

Second, inspired by work on model cards (Mitchell et al., 2019), we propose a standard *summarization data card* which covers the different items in our ontology as well as important properties. This can benefit future dataset developers in making their work more accessible, easy to discover and comparable against other datasets.

We hope that our work helps chart the way for future work in the field, as well enable researchers from various disciplines in exploring the wide range of existing summarization datasets.

To conclude we make the following contributions:

- We formulate an ontology for summarization datasets, covering various axes regrading the samples, their collection and distribution.
- We survey a wide range of 133 datasets covering datasets in 104 languages, showing that our ontology is rich enough to surface five important trends and challenges for future work, including the lack of terminology, the subpar quality of datasets in low-resource languages, and more.
- We make available two resources – a web interface to interact with our survey and discover summarization datasets, and a formal *summarization data card*, which will help streamline future datasets.

Shape	Example
Paragraph(s)	Canadian authorities say she was being sought for extradition to the US, where the company is being investigated for possible violation of sanctions against Iran. Canada's justice department said Meng was arrested in Vancouver on Dec. 1...
One Sentence	John Higgins believes he has the ability and the form to win his fifth World Championship.
Span	**Earlier this year the art world made a rare discovery – a painting by Leonardo da Vinci**. Only some 15 paintings by Leonardo still exist, including the "Mona Lisa". **So, this painting – known as "Salvator Mundi" or "Savior of the World" – is a truly remarkable find.**
Highlights	- Web site operator wants to publish encyclopedia about Harry Potter novels. - Judge awards "Harry Potter" author J.K. Rowling and publisher $6,750 in damages. - Rowling says she has long planned to publish her own encyclopedia.

Table 1: The different shapes of summaries in our collection. Examples are taken from (Cheng and Lapata, 2016; Hermann et al., 2015), (Lins et al., 2019), (Narayan et al., 2018) and (Fabbri et al., 2019).

2 Dataset Collection

Our survey covers 133 datasets collected over nearly 20 years, spanning from 2005 to 2024. The full list of references can be found in Appendix A.

Venues. Most datasets are drawn from papers published in top-tier NLP, ML and AI venues and workshops[2]. In order to expand the supported languages, we also include datasets published through shared tasks and in language specific journals and workshops.

Filtering. We manually verify that each of the datasets is available for use, and exclude papers describing currently inaccessible resources. We include papers that explicitly focused on creating summarization datasets or that focused on summarization models and produced data as a by-product.

Collected Data. For each dataset, we gather information on its properties, collection methods and distribution. This methodology allowed categorizing the current resources options for researchers, analyzing emerging trends and identifying the gaps in the field.

In order to analyse the current state of summarization datasets, we gathered additional information not covered in the ontology: the scale of data by number of samples, the levels of abstraction of the summaries and the data quality assessment methods applied to the collected data.

3 Ontology of Summarization Datasets

We propose a categorization of summarization datasets across seven orthogonal axes, and provide an ontology for each, depicted in Figure 2, and elaborated below. This ontology helps contextualize the extensive research in summarization and identifies missing types of datasets which could be addressed in follow up work.

3.1 Language and Language Modality

In the context of summarization, language choice entails the language used for input and output texts. We find a diverse set of datasets in 104 languages (see Appendix A).

In addition, we categorize each dataset into three categories of language modality:

Monolingual. These datasets contain texts in a single language for both input and output.

Multilingual. These datasets can be thought of as a few monolingual datasets together, as they contain document-summary pairs in various languages.

Cross-lingual. In these datasets, the language of the output summary is different from the language of the input document.

3.2 Domain

Domain is the area of knowledge from which the documents are sourced. We identified nine domains that encompass the vast majority of datasets: *News*, which relies on articles from online platforms; *Dialogue*, which includes sources such as meeting transcripts, movie scenes, interviews, and customer service interactions; *Scientific* papers, most commonly paired with their abstract as summary; *Legal* documents such as government bills and patent documents; *Encyclopedia* entries, primarily based on Wikipedia; *Opinions and arguments* including reviews, editorials, and debates; *Social media* texts,

[2]Venues: ACL, NAACL, EMNLP, LREC, EACL, SIGIR, AAAI, COLING, IEEE, NeurIPS, CoNLL, and TACL.

which are typically more freeform; *Literature* such as books and poetry; and *Instructional* content, like "how-to" guides.

3.3 Summary Shape

We categorize the shape of the summaries in the datasets into four broad types, shown in Table 1.

Single-sentence summary. These are summaries ranging from a few words (like titles) to a lengthy sentence, composed of several dozens of words. This shape is sometimes referred to as "extreme summarization" (Narayan et al., 2018) or "headline generation" (Napoles et al., 2012).

Highlights summary. These summaries are constructed from a few, possibly independent sentences. For example, they could be bullet points (e.g., from news articles).

Paragraph summary. These datasets contain output texts composed of one or more paragraphs of coherent free text. Examples of this shape include abstracts of scientific papers and sub-titles of news articles.

Span summary. Built to accommodate extractive summarization, these datasets use spans of text directly taken from the document as summaries.

3.4 Sources for Supervision and Annotation Efforts

We categorize datasets by how much manual annotation effort they required. On opposite ends are datasets that were gathered in an **Automatic** way, mainly through scraping web resources, and **human annotated** datasets — in which annotators wrote the gold summaries. We also find **semi-automatic** datasets — in which human annotators are involved in the process but only later in the pipeline. Datasets created using this method typically provide annotators with summary options that they can correct, filter, or accept.

We argue that an additional often overlooked dimension is the objective for which the summaries were originally created, marked as *Supervision* in Figure 2.

Natural. These datasets utilize summaries that were organically created. For example, highlights in news articles (Cheng and Lapata, 2016; Hermann et al., 2015) and educational summaries published along text to help students (Ladhak et al., 2020b).

Distant. Similarly to the Natural category, these datasets use pre-existing texts as annotations. However, these texts were not originally created with the intention of serving as summaries, such as abstracts of scientific papers (Cohan et al., 2018) or titles and subheadings of news articles (Napoles et al., 2012).

Dedicated. This category features datasets in which the summaries were written due to the creators' efforts and did not exist beforehand. It includes mostly human annotated datasets, but also synthetically created ones.

3.5 Availability

The distribution of datasets depends on how accessible they are to the research community. We identify three main forms of data availability.

Publicly Available. These datasets are published through a platform like Hugging Face.

URL-based reconstruction. In this category, the data is not directly published; instead, a method to reconstruct it is provided. For example, a list of URLs to scrape as done in (Grusky et al., 2018), or data preparation scripts as done in (Kryściński et al., 2021).

Upon request. This availability type, which is most restricting, usually requires filling a request form or contacting the authors in order to receive the data.

4 The State of Summarization Datasets

We thoroughly examine 133 summarization datasets under the lens of our ontology. Figure 2 presents the quantitative dataset distribution as well as example categorization of several datasets. Following, we discuss five major observations regarding the current state and challenges in summarization datasets.

4.1 There is a Lack in Common Terminology

While summarization is a versatile and expressive task applicable in many scenarios, we find that summarization datasets are presented using inconsistent terminology and often lacking details. This makes it difficult for researchers to identify available resources and to effectively compare and contrast them.

For example, determining the *shape* of the summaries (such as paragraph or highlights) often required to access the dataset and examine samples;

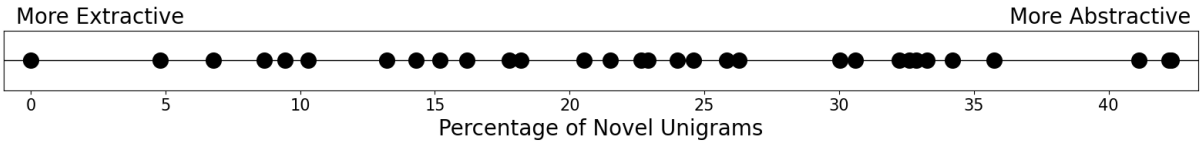

Figure 3: Every dot on the graph corresponds to a dataset in our collection. Each dot's position indicates the percentage of unique uni-grams in the summaries that are not found in the source texts. Only a minority of the datasets report evaluation of the properties of the summaries.

understanding how the summaries were created frequently involved visiting the original website from which the data was scraped; while datasets often do not mention the language in which they were annotated, in contrast with recent emphasis on explicit language disclosure (Bender, 2019).

All of these highlight the necessity of a shared ontology expressive enough to capture currently important aspects in summarization datasets, and guide dataset developers in reporting their work.

Abstractive vs. extractive summarization is no longer a useful distinction. An important case where terminology is lacking is the distinction between *extractive* summarization, where models are required to select sentences from the source document and *abstractive* summarization, where generative models can also produce novel text.

With the advent of powerful generative LLMs, our survey indicates that these distinctions are no longer useful, with many of the surveyed datasets not specifying whether their annotations were either extractive or abstractive.

Instead, we argue that abstractiveness is a spectrum rather than a binary trait. To support this, in Figure 3 we plot the novel n-grams ratio for datasets in our collection. This metric, which we find to be the most commonly used for measuring levels of abstractions, measures the percentage of n-grams appearing in the summaries but not in the document. Formally (Narayan et al., 2018):

$$\frac{\text{\#New n-grams}}{\text{\#n-grams}} \quad (1)$$

On this scale, 0% can be achieved by restricting the summary to contain only vocabulary from the source document and thus lean towards extractive. While at the other end of the scale, 40% of the 1-grams in the summaries were novel and thus leaning towards abstractive. Other less common metrics include Coverage and Density (Grusky et al., 2018) as well as Abstractivity (ABS; Bommasani and Cardie, 2020). All of which are based on the overlap between the summary and the source document.

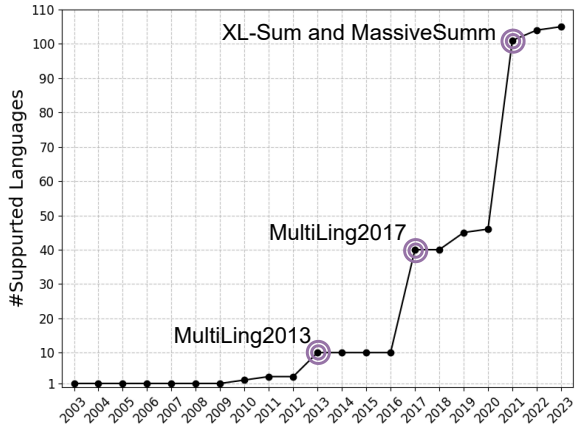

Figure 4: The rise in the number of languages supported by summarization resources was achieved mainly through multilingual datasets. Most languages in our survey are only supported through them.

4.2 News is the Most Popular Domain, Despite its Major Shortcomings

Roughly 50% of the datasets we surveyed use news articles as source documents. This popularity may be due to the large number of online news platforms in various languages that allow for easy scraping. Moreover, many of these articles already include human-written summaries or distant supervision, in the form of highlights, titles, and subtitles. We argue that this wide availability and relative ease of obtaining distant supervision is in fact a *double-edged sword*, harming the quality of the gold summaries, while presenting a rather simple challenge for summarization models.

First, the news domain may present a simple instance of the summarization task. News articles typically follow "inverted pyramid writing style" — i.e., the most important information answering the 5 w's (who? what? etc.) appears in the first few sentences (Pottker, 2003). Consequently, a summarization model which trivially outputs the first sentences of the news article can often achieve

good results (Cohen et al., 2021).

Second, news summarization datasets are prone to be of low quality, because they often rely on automatic heuristics. Tejaswin et al. (2021) manually evaluated two datasets with one sentence summaries, one which uses the title as summary (Rush et al., 2015), and another which uses a bold sentence in the beginning of the article (Narayan et al., 2018). They found that about half of the samples are either incomplete or contain concepts or entities that do not appear in the rest of the document, and hence often cannot be deduced from it.

Beyond the news domain, the scientific domain, characterized by longer texts and complex terminology is popular (Cohan et al., 2018), yet it remains predominantly English-centric. Official documents, such as those in the legal domain, patents, and government bills (e.g., Sharma et al., 2019; Kornilova and Eidelman, 2019; Huang et al., 2021), share similar complexity in that salient information is usually distributed throughout the document, and are more multilingual. We also find that dialogues are increasingly attracting the attention of the research community, with this domain being the second most popular, and that Wikipedia articles are a valuable resource due to their guidelines and API, as well as for cross-lingual research because they cover the same topics in multiple languages.

4.3 Datasets for Low-Resource Languages are Often Low Quality and Hard to Discover

The attention to performing NLP research on a diverse set of languages has been increasing in recent years (Joshi et al., 2020; Ruder, 2022). Evidently, while English is by far the most popular language in our survey, we cover resources for 104 different languages (see Appendix A for the complete list). We find that this encouraging versatility was achieved mainly through multi-lingual datasets, as 81 of the languages surveyed are supported solely through them. In Figure 4, we illustrate the growth in the number of languages with available resources over the years, emphasizing the dramatic rise following the introduction of multilingual resources.

However, while providing important diversity, we identify limitations in how multilingual summarization datasets are created and reported.

Discovering existing datasets in low-resource languages proves challenging. Given that the majority of languages in our survey are available only within large multilingual corpora, researchers must comb through all multilingual datasets in search for their language. This is challenging as there is currently no unified platform that houses all available datasets and allows to search by the desired language.

To make matters worse, determining which languages are present in each multilingual dataset proved challenging, as we mentioned before, datasets often do not explicitly list the supported languages in the paper.

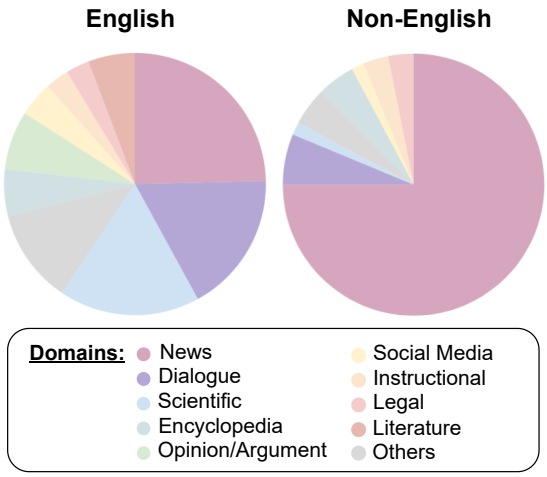

Figure 5: The distribution of domains for English datasets versus non-English datasets. While English domains are diverse, our survey shows most datasets for other languages are comprised of news articles.

Datasets in low-resource languages are less diverse in terms of domain, and of poorer quality. We find that the over-reliance on the news domain is considerably worse in non-English datasets. In Figure 5, we compare the domain distribution between English and non-English resources. We find that more than 70% of datasets for these languages are from the news domain. This exacerbates the problems we observed regarding the news domain. Indeed, Urlana et al. (2022) reported that 97% of the Telugu samples in XL-SUM (Hasan et al., 2021) and MassiveSum (Varab and Schluter, 2021) were of low quality, with similar results in Hindi, Gujarati and Marathi.

4.4 Most Annotations were not Intended to be Used as Summaries

We found only 35 *dedicated* datasets, i.e., where the annotations were created specifically for the task and did not exist before. This is three times fewer than datasets that include existing annotations.

Not only that the vast majority of annotations

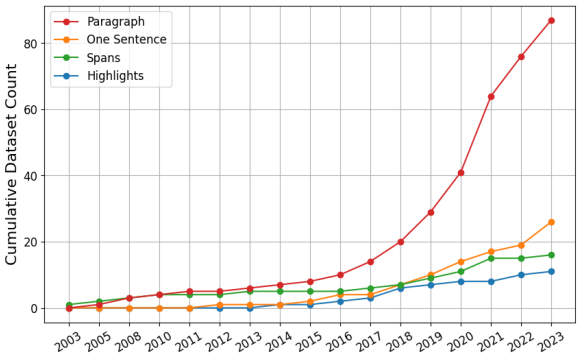

Figure 6: There is a clear trend of collecting summaries of paragraph shape. We note some datasets offer a few shapes.

available were gathered automatically, but datasets collected this way are an order of magnitude bigger: human annotated datasets were on average more than 10 times smaller in number of samples than automatically collected datasets and thus allow for training big models.

We categorized 43% of the datasets as *Distant* supervision, meaning the human authors did not explicitly intend to write a summary of the source document. This includes proxies of summaries like titles and subtitles in news articles, abstracts of scientific papers or descriptions of videos or interviews. Other approaches we consider as distant are social media metadata as summaries, as introduced in Newsroom (Grusky et al., 2018) and beginning with summaries and searching for longer texts to support them (Ghalandari et al., 2020; Saxena and Keller, 2024).

We also find a mismatch between the datasets we create and the ones we use. In Figure 6, we show a clear trend over the years of collecting the *paragraph* shape. This might be considered a natural shape for summarization as it is also the most common for collecting data through human annotation in our collection. This comes in contrast to Highlights being very commonly used in research. This is evident from CNN-Daily Mail, perhaps the most popular summarization dataset at the time of writing, which uses this shape.

4.5 Copyright Concerns Limit the Distribution of Datasets

Our survey shows that there is a recent trend away from publishing datasets in a public manner, instead making them available through reconstruction scripts from URLs (Grusky et al., 2018; Varab and Schluter, 2020; Kryściński et al., 2021), which may be due to growing legal concerns (Karamolegkou et al., 2023).

Reconstruction may be a temporary solution, which does not address the underlying legal issues. Moreover, this practice may lead to inconsistencies between local instances of the same dataset obtained at different times, due to changes in URL contents.

Recently, even datasets which were previously publicly available are being called into question. For example, the popular CNN/Daily Mail and XSUM are considered to be "legally questionable" (Wang et al., 2022a). In a similar vein, the New York Times Corpus (Sandhaus, 2008), despite being published by the newspaper itself, was recently withdrawn and is no longer available. This may be another downside of the field's over-reliance on the news domain.

5 Standardizing Summarization Datasets

Our survey surfaces a lack in terminology and reporting of the vast number of summarization datasets. This challenges both users of datasets, who may find it hard knowing whether there exists a dataset specifically tailored to their needs (e.g., in terms of language, domain, or shape), as well as researchers in the field, looking to identify current bottlenecks and directions for future research. Here we outline two contributions we make towards making datasets more standard and accessible.

Interactive web interface. We will make available a web UI populated with our ontology of 133 summarization datasets, along with our annotations of the seven axes which characterize each dataset (See Figure 7). Users of this interface can search for a dataset by specifying the criteria they are looking for. We will welcome community contributions to add new datasets, or past datasets not already included in it. We hope that this interface can beacome a central hub, allowing users to easily find relevant datasets, without the need to manually sift through hundreds of papers.

Template for a summarization *data card*. In recent years, model and data cards have helped set agreed reporting standards, formalizing what details are reported when publishing a new resource (Mitchell et al., 2019). Moving forward, we suggest that summarization research adopt a similar approach. In Table 2 we propose a tem-

Summarization Data Card
Sample information:
Languages:
List all supported languages
Summary Shape:
Paragraph/One Sentence/Highlights/Span
Domain:
Example: News/Scientific/Dialogues/etc.
Size:
Number of document-summary pairs
Annotation information:
Annotation efforts:
Automatic, Human annotations, Semi-automatic
Source of supervision:
Natural (summaries created organically)/
Distant (annotations are proxies of summaries)/
Dedicated (annotations created by researchers)
Brief description of the summaries' source:
Example: digests of legal documents
Data quality assessment:
Abstraction level:
1-to-4-gram ratios
Compression rate: $\frac{\text{doc length (\#words)}}{\text{summary length (\#words)}}$
Human evaluation: *Yes/No*
Availability details:
How is the data made accessible:
Publicly Available /
URL-based Reconstruction /
Upon Request
Copyrights information:
License

Table 2: Template for a summarization data card.

plate for a summarization data card which follows findings from our survey.

Specifically, we focus on four key elements to report: (1) sample information: shape, languages, domain; (2) annotation information, indicating how were the summaries created: e.g., via scraping or manual effort; (3) data quality assessment: we propose to report the level of abstraction via n-gram novelty, the Compression rate (Grusky et al., 2018), as well as human evaluation of the data, if it was performed; and (4) availability: indicating copyright licensing and download information.

6 Related Work

Several surveys have explored different aspects of automatic summarization (Kryściński et al., 2019; Cajueiro et al., 2023; Sharma and Sharma, 2022), yet their focus was more on models rather than datasets. To the best of our knowledge, none of the previous work has focused on the state of low-resource languages, nor looked at the dataset collection method and distribution. (Dernoncourt et al., 2018) published a repository of available corpora for summarization. However, it included only a limited number of datasets (21), with just four that are not in English.

Several surveys focus on a sub-task of summarization: (Koh et al., 2022) reviewed long document summarization; (Jangra et al., 2023) studied multimodal summarization; and (Tuggener et al., 2021) conducted a survey of dialogue summarization datasets.

A few papers focused on assessing the properties of popular datasets. Bommasani and Cardie (2020) conducted evaluation of 10 famous summarization datasets using novel metrics and Tejaswin et al. (2021) manually analyse 600 samples from three datasets.

7 Conclusion

We carried out an extensive survey of 133 summarization datasets in 104 languages, introducing a new ontology covering sample information, collection methods and data distribution. We offer a detailed report on the current state of resources in the field, surfacing challenges due to the lack of terminology, as well as findings on low resource languages, data quality and availability. To support researchers, we release two resources: a platform to search available datasets by our categorization and a template for a summarization data card, to help streamline future datasets.

8 Limitations

Although this survey provides a thorough overview of the available summarization datasets, we wish to acknowledge several limitations. First, despite our best efforts, we may not have found all available resources. This is especially true for low resource languages and datasets published by less-established conferences.

Second, we decided to omit from our ontology some axis. Specifically, we did not analysed data about sub-tasks of summarizations such as multimodal, long-form or aspect based. We note that the available resources are still included in our platform and marked accordingly.

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234.

Abdulaziz Alhamadani, Xuchao Zhang, Jianfeng He, and Chang-Tien Lu. 2022. Lans: Large-scale arabic news summarization corpus. *arXiv preprint arXiv:2210.13600*.

Diego Antognini and Boi Faltings. 2020. Gamewikisum: a novel large multi-document summarization dataset. *arXiv preprint arXiv:2002.06851*.

Dennis Aumiller, Ashish Chouhan, and Michael Gertz. 2022. Eur-lex-sum: A multi-and cross-lingual dataset for long-form summarization in the legal domain. *arXiv preprint arXiv:2210.13448*.

Soran Badawi. 2023. Kurdsum: A new benchmark dataset for the kurdish text summarization. *Natural Language Processing Journal*, 5:100043.

Botond Barta, Dorina Lakatos, Attila Nagy, Milán Konor Nyist, and Judit Ács. 2024. From news to summaries: Building a hungarian corpus for extractive and abstractive summarization. *arXiv preprint arXiv:2404.03555*.

Emily Bender. 2019. The benderrule: On naming the languages we study and why it matters. *The Gradient*.

Sumit Bhatia, Prakhar Biyani, and Prasenjit Mitra. 2014. Summarizing online forum discussions–can dialog acts of individual messages help? In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 2127–2131.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2021. Crosssum: Beyond english-centric cross-lingual summarization for 1,500+ language pairs. *arXiv preprint arXiv:2112.08804*.

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.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S Weld. 2020. Tldr: Extreme summarization of scientific documents. *arXiv preprint arXiv:2004.15011*.

Daniel O Cajueiro, Arthur G Nery, Igor Tavares, Maísa K De Melo, Silvia A dos Reis, Li Weigang, and Victor RR Celestino. 2023. A comprehensive review of automatic text summarization techniques: method, data, evaluation and coding. *arXiv preprint arXiv:2301.03403*.

Yue Cao, Xiaojun Wan, Jinge Yao, and Dian Yu. 2020. Multisumm: towards a unified model for multilingual abstractive summarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11–18.

Ziqiang Cao, Chengyao Chen, Wenjie Li, Sujian Li, Furu Wei, and Ming Zhou. 2016. Tgsum: Build tweet guided multi-document summarization dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Paula CF Cardoso, Erick G Maziero, Mara Luca Castro Jorge, Eloize MR Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Gracas Volpe Nunes, and Thiago AS Pardo. 2011. Cstnews-a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105.

Jingqiang Chen and Hai Zhuge. 2018. Abstractive text-image summarization using multi-modal attentional hierarchical rnn. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4046–4056.

Kai Chen, Guanyu Fu, Qingcai Chen, and Baotian Hu. 2021a. A large-scale chinese long-text extractive summarization corpus. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7828–7832. IEEE.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2021b. Summscreen: A dataset for abstractive screenplay summarization. *arXiv preprint arXiv:2104.07091*.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021c. Dialogsum: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762*.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252*.

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.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*.

Nachshon Cohen, Oren Kalinsky, Yftah Ziser, and Alessandro Moschitti. 2021. Wikisum: Coherent summarization dataset for efficient human-evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the*

11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 212–219.

Ed Collins, Isabelle Augenstein, and Sebastian Riedel. 2017. A supervised approach to extractive summarisation of scientific papers. *arXiv preprint arXiv:1706.03946*.

Debtanu Datta, Shubham Soni, Rajdeep Mukherjee, and Saptarshi Ghosh. 2023. Mildsum: A novel benchmark dataset for multilingual summarization of indian legal case judgments. *arXiv preprint arXiv:2310.18600*.

Franck Dernoncourt, Mohammad Ghassemi, and Walter Chang. 2018. A repository of corpora for summarization. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Moussa Kamal Eddine, Antoine J-P Tixier, and Michalis Vazirgiannis. 2020. Barthez: a skilled pretrained french sequence-to-sequence model. *arXiv preprint arXiv:2010.12321*.

Mahmoud El-Haj and Rim Koulali. 2013. Kalimat a multipurpose arabic corpus. *Culture*, 2:1–359.

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

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Ali Faheem, Faizad Ullah, Muhammad Sohaib Ayub, and Asim Karim. 2024. Urdumasd: A multimodal abstractive summarization dataset for urdu. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17245–17253.

Mehrdad Farahani, Mohammad Gharachorloo, and Mohammad Manthouri. 2021. Leveraging parsbert and pretrained mt5 for persian abstractive text summarization. In *2021 26th International computer conference, computer society of Iran (CSICC)*, pages 1–6. IEEE.

Mehwish Fatima and Michael Strube. 2021. A novel wikipedia based dataset for monolingual and cross-lingual summarization. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 39–50.

Guy Feigenblat, Chulaka Gunasekara, Benjamin Sznajder, Sachindra Joshi, David Konopnicki, and Ranit Aharonov. 2021. Tweetsumm–a dialog summarization dataset for customer service. *arXiv preprint arXiv:2111.11894*.

Svitlana Galeshchuk. 2023. Abstractive summarization for the ukrainian language: Multi-task learning with hromadske. ua news dataset. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 49–53.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the wikipedia current events portal. *arXiv preprint arXiv:2005.10070*.

George Giannakopoulos, John Conroy, Jeff Kubina, Peter A Rankel, Elena Lloret, Josef Steinberger, Marina Litvak, and Benoit Favre. 2017. Multiling 2017 overview. In *Proceedings of the MultiLing 2017 workshop on summarization and summary evaluation across source types and genres*, pages 1–6.

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.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Association for Computational Linguistics.

Philip John Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1066–1076. Association for Computational Linguistics.

Pritom Jyoti Goutom, Nomi Baruah, and Paramananda Sonowal. 2023. An abstractive text summarization using deep learning in assamese. *International Journal of Information Technology*, 15(5):2365–2372.

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*.

Vivek Gupta, Prerna Bharti, Pegah Nokhiz, and Harish Karnick. 2021. Sumpubmed: Summarization dataset of pubmed scientific articles. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 292–303.

Ilya Gusev. 2020. Dataset for automatic summarization of russian news. In *Artificial Intelligence and Natural Language: 9th Conference, AINL 2020, Helsinki, Finland, October 7–9, 2020, Proceedings 9*, pages 122–134. Springer.

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*.

Laura Hasler, Constantin Orasan, and Ruslan Mitkov. 2003. Building better corpora for summarisation. In *Proceedings of corpus linguistics*, pages 309–319.

Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. Wikiasp: A dataset for multi-domain aspect-based summarization. *Transactions of the Association for Computational Linguistics*, 9:211–225.

Hiroaki Hayashi, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. What's new? summarizing contributions in scientific literature. *arXiv preprint arXiv:2011.03161*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. Lcsts: A large scale chinese short text summarization dataset. *arXiv preprint arXiv:1506.05865*.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. *arXiv preprint arXiv:2104.02112*.

Anubhav Jangra, Sourajit Mukherjee, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. 2023. A survey on multi-modal summarization. *ACM Computing Surveys*, 55(13s):1–36.

Xiang Jiang and Markus Dreyer. 2024. Ccsum: A large-scale and high-quality dataset for abstractive news summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7299–7329.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright violations and large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7412, Singapore. Association for Computational Linguistics.

Tannon Kew, Marek Kostrzewa, and Sarah Ebling. 2023. 20 minuten: A multi-task news summarisation dataset for german. In *Proceedings of the Swiss Text Analytics Conference*, 8th edit., pages 1–13. Association for Computational Linguistics.

Misha Khalman, Yao Zhao, and Mohammad Saleh. 2021. Forumsum: A multi-speaker conversation summarization dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4592–4599.

Alvi Khan, Fida Kamal, Mohammad Abrar Chowdhury, Tasnim Ahmed, Md Tahmid Rahman Laskar, and Sabbir Ahmed. 2023. Banglachq-summ: An abstractive summarization dataset for medical queries in bangla conversational speech. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 85–93.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2018. Abstractive summarization of reddit posts with multi-level memory networks. *arXiv preprint arXiv:1811.00783*.

Svea Klaus, Ria Van Hecke, Kaweh Djafari Naini, Ismail Sengor Altingovde, Juan Bernabé-Moreno, and Enrique Herrera-Viedma. 2022. Summarizing legal regulatory documents using transformers. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2426–2430.

Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM computing surveys*, 55(8):1–35.

Anastassia Kornilova and Vlad Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation. *arXiv preprint arXiv:1910.00523*.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2020. Liputan6: A large-scale indonesian dataset for text summarization. *arXiv preprint arXiv:2011.00679*.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The ami meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*, pages 1–4.

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2021. Booksum: A collection of datasets for long-form narrative summarization. *arXiv preprint arXiv:2105.08209*.

Kemal Kurniawan and Samuel Louvan. 2018. Indosum: A new benchmark dataset for indonesian text summarization. In *2018 International Conference on Asian Language Processing (IALP)*, pages 215–220. IEEE.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020a. Wikilingua: A new benchmark dataset for cross-lingual abstractive summarization. *arXiv preprint arXiv:2010.03093*.

Faisal Ladhak, Bryan Li, Yaser Al-Onaizan, and Kathleen McKeown. 2020b. Exploring content selection in summarization of novel chapters. *arXiv preprint arXiv:2005.01840*.

Nicola Landro, Ignazio Gallo, Riccardo La Grassa, and Edoardo Federici. 2022. Two new datasets for italian-language abstractive text summarization. *Information*, 13(5):228.

Guy Lev, Michal Shmueli-Scheuer, Jonathan Herzig, Achiya Jerbi, and David Konopnicki. 2019. Talksumm: A dataset and scalable annotation method for scientific paper summarization based on conference talks. *arXiv preprint arXiv:1906.01351*.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, Chengqing Zong, et al. 2018. Multi-modal sentence summarization with modality attention and image filtering. In *IJCAI*, pages 4152–4158.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102.

Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020. Vmsmo: Learning to generate multimodal summary for video-based news articles. *arXiv preprint arXiv:2010.05406*.

Yunlong Liang, Fandong Meng, Jinan Xu, Jiaan Wang, Yufeng Chen, and Jie Zhou. 2022. Summary-oriented vision modeling for multimodal abstractive summarization. *arXiv preprint arXiv:2212.07672*.

Haitao Lin, Liqun Ma, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021. Csds: A fine-grained chinese dataset for customer service dialogue summarization. *arXiv preprint arXiv:2108.13139*.

Rafael Dueire Lins, Hilario Oliveira, Luciano Cabral, Jamilson Batista, Bruno Tenorio, Rafael Ferreira, Rinaldo Lima, Gabriel de França Pereira e Silva, and Steven J Simske. 2019. The cnn-corpus: A large textual corpus for single-document extractive summarization. In *Proceedings of the ACM Symposium on Document Engineering 2019*, pages 1–10.

Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2023. Generating a structured summary of numerous academic papers: Dataset and method. *arXiv preprint arXiv:2302.04580*.

Xiaojun Liu, Shunan Zang, Chuang Zhang, Xiaojun Chen, and Yangyang Ding. 2022. Clts+: a new chinese long text summarization dataset with abstractive summaries. In *International Conference on Artificial Neural Networks*, pages 73–84. Springer.

Xiaojun Liu, Chuang Zhang, Xiaojun Chen, Yanan Cao, and Jinpeng Li. 2020. Clts: a new chinese long text summarization dataset. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 531–542. Springer.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. *arXiv preprint arXiv:2010.14235*.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Ridwan Mahbub, Ifrad Khan, Samiha Anuva, Md Shihab Shahriar, Md Tahmid Rahman Laskar, and Sabbir Ahmed. 2023. Unveiling the essence of poetry: Introducing a comprehensive dataset and benchmark for poem summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14878–14886.

Valentin Malykh, Konstantin Chernis, Ekaterina Artemova, and Irina Piontkovskaya. 2020. Sumtitles: a summarization dataset with low extractiveness. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5718–5730.

Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. 2021. Bringing structure into summaries: a faceted summarization dataset for long scientific documents. *arXiv preprint arXiv:2106.00130*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Itai Mondshine, Tzuf Paz-Argaman, Asaf Achi Mordechai, and Reut Tsarfaty. 2024. Hesum: a novel dataset for abstractive text summarization in hebrew. In *Proceedings of the The Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 26–36.

Courtney Napoles, Matthew R Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the joint workshop on automatic knowledge base construction and web-scale knowledge extraction (AKBC-WEKEX)*, pages 95–100.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the

summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Khanh Nguyen and Hal Daumé III. 2019. Global voices: Crossing borders in automatic news summarization. *arXiv preprint arXiv:1910.00421*.

Van-Hau Nguyen, Thanh-Chinh Nguyen, Minh-Tien Nguyen, and Nguyen Xuan Hoai. 2019. Vnds: A vietnamese dataset for summarization. In *2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 375–380. IEEE.

Viktoria Ondrejova and Marek Šuppa. 2024. Slovaksum: A large scale slovak summarization dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14916–14922.

Pedro Henrique Paiola, Gabriel Lino Garcia, Danilo Samuel Jodas, João Vitor Mariano Correia, Luis Afonso Sugi, and João Paulo Papa. 2024. Recognasumm: A novel brazilian summarization dataset. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 575–579.

C Palen-Michel and C Lignos. Lr-sum: Summarization for less-resourced languages (2022).

Min Su Park and Eunil Park. 2023. Acl ta-da: A dataset for text summarization and generation. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, pages 1233–1239.

Laura Perez-Beltrachini and Mirella Lapata. 2022. Models and datasets for cross-lingual summarisation. *arXiv preprint arXiv:2202.09583*.

Horst Pottker. 2003. News and its communicative quality: the inverted pyramid—when and why did it appear? *Journalism Studies*, 4(4):501–511.

Allen Roush and Arvind Balaji. 2020. Debatesum: A large-scale argument mining and summarization dataset. *arXiv preprint arXiv:2011.07251*.

Sebastian Ruder. 2022. The State of Multilingual AI. http://ruder.io/state-of-multilingual-ai/.

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, Lisbon, Portugal. Association for Computational Linguistics.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*.

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

Shrey Satapara, Bhavan Modha, Sandip Modha, and Parth Mehta. 2022. Fire 2022 ilsum track: Indian language summarization. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 8–11.

Rohit Saxena and Frank Keller. 2024. Select and summarize: Scene saliency for movie script summarization. *arXiv preprint arXiv:2404.03561*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. Mlsum: The multilingual summarization corpus. *arXiv preprint arXiv:2004.14900*.

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

Grishma Sharma and Deepak Sharma. 2022. Automatic text summarization methods: A comprehensive review. *SN Computer Science*, 4(1):33.

Xiaorui Shi. 2023. Mcls: A large-scale multimodal cross-lingual summarization dataset. In *China national conference on chinese computational linguistics*, pages 273–288. Springer.

Encarnación Segarra Soriano, Vicent Ahuir, Lluís-F Hurtado, and José González. 2022. Dacsa: A large-scale dataset for automatic summarization of catalan and spanish newspaper articles. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5931–5943.

Sajad Sotudeh, Hanieh Deilamsalehy, Franck Dernoncourt, and Nazli Goharian. 2021. Tldr9+: a large scale resource for extreme summarization of social media posts. *arXiv preprint arXiv:2110.01159*.

Milan Straka, Nikita Mediankin, Tom Kocmi, Zdeněk Žabokrtskỳ, Vojtěch Hudeček, and Jan Hajic. 2018. Sumeczech: Large czech news-based summarization dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Marek Šuppa and Jergus Adamec. 2020. A summarization dataset of slovak news articles. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6725–6730.

Shahbaz Syed, Roxanne El Baff, Johannes Kiesel, Khalid Al Khatib, Benno Stein, and Martin Potthast. 2020. News editorials: Towards summarizing long argumentative texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5384–5396.

Hoang Thang Ta, Abu Bakar Siddiqur Rahman, Navonil Majumder, Amir Hussain, Lotfollah Najjar, Newton Howard, Soujanya Poria, and Alexander Gelbukh. 2023. Wikides: A wikipedia-based dataset for generating short descriptions from paragraphs. *Information Fusion*, 90:265–282.

Sotaro Takeshita, Tommaso Green, Niklas Friedrich, Kai Eckert, and Simone Paolo Ponzetto. 2024a. Cross-lingual extreme summarization of scholarly documents. *International Journal on Digital Libraries*, 25(2):249–271.

Sotaro Takeshita, Tommaso Green, Ines Reinig, Kai Eckert, and Simone Paolo Ponzetto. 2024b. Aclsum: A new dataset for aspect-based summarization of scientific publications. *arXiv preprint arXiv:2403.05303*.

Priyam Tejaswin, Dhruv Naik, and Pengfei Liu. 2021. How well do you know your summarization datasets? *arXiv preprint arXiv:2106.11388*.

Pavel Tikhonov and Valentin Malykh. 2022. Wikimulti: a corpus for cross-lingual summarization. In *Conference on Artificial Intelligence and Natural Language*, pages 60–69. Springer.

Nhi-Thao Tran, Minh-Quoc Nghiem, Nhung TH Nguyen, Ngan Luu-Thuy Nguyen, Nam Van Chi, and Dien Dinh. 2020. Vims: a high-quality vietnamese dataset for abstractive multi-document summarization. *Language Resources and Evaluation*, 54(4):893–920.

Don Tuggener, Margot Mieskes, Jan Milan Deriu, and Mark Cieliebak. 2021. Are we summarizing the right way?: a survey of dialogue summarization data sets. In *Conference on Empirical Methods in Natural Language Processing (EMNLP), Punta Cana, Dominican Republic (online), 7-11 November 2021*, pages 107–118. Association for Computational Linguistics.

Jan Ulrich, Gabriel Murray, and Giuseppe Carenini. 2008. A publicly available annotated corpus for supervised email summarization. In *AAAI08 email workshop*, pages 1–6.

Ashok Urlana, Pinzhen Chen, Zheng Zhao, Shay B Cohen, Manish Shrivastava, and Barry Haddow. 2023. Pmindiasum: Multilingual and cross-lingual headline summarization for languages in india. *arXiv preprint arXiv:2305.08828*.

Ashok Urlana, Nirmal Surange, Pavan Baswani, Priyanka Ravva, and Manish Shrivastava. 2022. Tesum: Human-generated abstractive summarization corpus for telugu. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5712–5722.

Daniel Varab and Natalie Schluter. 2020. Danewsroom: A large-scale danish summarisation dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6731–6739.

Daniel Varab and Natalie Schluter. 2021. Massivesumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161.

Yash Verma, Anubhav Jangra, Raghvendra Kumar, and Sriparna Saha. 2023. Large scale multi-lingual multi-modal summarization dataset. *arXiv preprint arXiv:2302.06560*.

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.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R Bowman. 2022a. Squality: Building a long-document summarization dataset the hard way. *arXiv preprint arXiv:2205.11465*.

Danqing Wang, Jiaze Chen, Hao Zhou, Xipeng Qiu, and Lei Li. 2021. Contrastive aligned joint learning for multilingual summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2739–2750.

Jiaan Wang, Fandong Meng, Ziyao Lu, Duo Zheng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022b. Clidsum: A benchmark dataset for cross-lingual dialogue summarization. *arXiv preprint arXiv:2202.05599*.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. *arXiv preprint arXiv:1606.02785*.

Zhongyu Wei and Wei Gao. 2018. Utilizing microblogs for automatic news highlights extraction. In *Social Media Content Analysis: Natural Language Processing and Beyond*, pages 277–296. World Scientific.

Han Wu, Mingjie Zhan, Haochen Tan, Zhaohui Hou, Ding Liang, and Linqi Song. 2023. Vcsum: A versatile chinese meeting summarization dataset. *arXiv preprint arXiv:2305.05280*.

Shi Xiaorui. 2023. Mcls: A large-scale multimodal cross-lingual summarization dataset. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 862–874.

Xinnuo Xu, Ondřej Dušek, Shashi Narayan, Verena Rieser, and Ioannis Konstas. 2021. Miranews: Dataset and benchmarks for multi-resource-assisted news summarization. *arXiv preprint arXiv:2109.10650*.

Nishant Yadav, Matteo Brucato, Anna Fariha, Oscar Youngquist, Julian Killingback, Alexandra Meliou, and Peter Haas. 2021. Subsume: A dataset for subjective summary extraction from wikipedia documents. In *New Frontiers in Summarization workshop (at EMNLP 2021)*.

Shweta Yadav, Deepak Gupta, and Dina Demner-Fushman. 2022. Chq-summ: A dataset for consumer healthcare question summarization. *arXiv preprint arXiv:2206.06581*.

Takashi Yamamura and Kazutaka Shimada. 2018. Annotation and analysis of extractive summaries for the kyutech corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R. Fabbri, Irene Li, Dan Friedman, and Dragomir R. Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. *Preprint*, arXiv:1909.01716.

Shunan Zang, Chuang Zhang, Xiaojun Liu, Xiaojun Chen, Peng Zhang, and Jie Liu. 2022. Newsfarm: A large-scale chinese corpus of long news summarization. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2260–2272. IEEE.

Jianmin Zhang and Xiaojun Wan. 2017. Towards automatic construction of news overview articles by news synthesis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2111–2116.

Rui Zhang and Joel Tetreault. 2019. This email could save your life: Introducing the task of email subject line generation. *arXiv preprint arXiv:1906.03497*.

Mingjun Zhao, Shengli Yan, Bang Liu, Xinwang Zhong, Qian Hao, Haolan Chen, Di Niu, Bowei Long, and Weidong Guo. 2021. Qbsum: A large-scale query-based document summarization dataset from real-world applications. *Computer Speech & Language*, 66:101166.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. *arXiv preprint arXiv:2104.05938*.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. Mediasum: A large-scale media interview dataset for dialogue summarization. *Preprint*, arXiv:2103.06410.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. Msmo: Multimodal summarization with multimodal output. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4154–4164.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. Ncls: Neural cross-lingual summarization. *arXiv preprint arXiv:1909.00156*.

Markus Zopf. 2018. Auto-hmds: Automatic construction of a large heterogeneous multilingual multi-document summarization corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Markus Zopf, Maxime Peyrard, and Judith Eckle-Kohler. 2016. The next step for multi-document summarization: A heterogeneous multi-genre corpus built with a novel construction approach. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1535–1545.

A Datasets Categorized by Language

In this section we list all datasets included in this survey per language.

Afrikaans: (Varab and Schluter, 2021), (Giannakopoulos et al., 2017)

Albanian: (Varab and Schluter, 2021), (Palen-Michel and Lignos)

Amharic: (Hasan et al., 2021), (Palen-Michel and Lignos), (Bhattacharjee et al., 2021), (Shi, 2023), (Varab and Schluter, 2021), (Liang et al., 2022), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Arabic: (Giannakopoulos et al., 2015), (El-Haj et al., 2010), (Hasan et al., 2021), (Nguyen and Daumé III, 2019), (El-Haj and Koulali, 2013), (Alhamadani et al., 2022), (Chouigui et al., 2021), (Varab and Schluter, 2021), (Liang et al., 2022), (Giannakopoulos et al., 2017), (Ladhak et al., 2020a), (Giannakopoulos et al., 2017), (Bhattacharjee et al., 2021), (Tikhonov and Malykh, 2022), (Xiaorui, 2023)

Armenian: (Palen-Michel and Lignos), (Varab and Schluter, 2021)

Assamese: (Urlana et al., 2023), (Goutom et al., 2023), (Varab and Schluter, 2021)

Aymara: (Varab and Schluter, 2021)

Azerbaijani: (Hasan et al., 2021), (Palen-Michel and Lignos), (Varab and Schluter, 2021), (Liang et al., 2022), (Giannakopoulos et al., 2017),(Bhattacharjee et al., 2021), (Xiaorui, 2023)

Bambara: (Varab and Schluter, 2021)

Basque: (Giannakopoulos et al., 2017)

Figure 7: Our interface allows researchers to determine if their specific needs are met by available resources. It is especially important for low resource languages.

Bengali: (Hasan et al., 2021), (Urlana et al., 2023), (Palen-Michel and Lignos), (Nguyen and Daumé III, 2019), (Khan et al., 2023), (Varab and Schluter, 2021), (Liang et al., 2022), (Verma et al., 2023), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Bislama: (Varab and Schluter, 2021)

Bosnian: (Palen-Michel and Lignos), (Varab and Schluter, 2021), (Cao et al., 2020), (Giannakopoulos et al., 2017)

Bulgarian: (Varab and Schluter, 2021), (Giannakopoulos et al., 2017), (Aumiller et al., 2022)

Burmese: (Hasan et al., 2021), (Palen-Michel and Lignos), (Varab and Schluter, 2021), (Liang et al., 2022), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Catalan: (Soriano et al., 2022), (Varab and Schluter, 2021), (Giannakopoulos et al., 2017)

Central Khmer: (Palen-Michel and Lignos), (Varab and Schluter, 2021)

Chinese: [3] (Giannakopoulos et al., 2015), (Hu et al., 2015), (Hasan et al., 2021), (Palen-Michel and Lignos), (Wu et al., 2023), (Liu et al., 2020), (Liu et al., 2022), (Zang et al., 2022), (Chen et al., 2021a), (Zhao et al., 2021), (Varab and Schluter, 2021), (Li et al., 2017), (Liang et al., 2022), (Lin et al., 2021), (Verma et al., 2023), (Li et al., 2020), (Wang et al., 2021), (Giannakopoulos et al., 2017), (Ladhak et al., 2020a), (Zhu et al., 2019), (Wang et al., 2022b), (Bhattacharjee et al., 2021), (Takeshita et al., 2024a), (Tikhonov and Malykh, 2022), (Xiaorui, 2023)

Croatian: (Varab and Schluter, 2021), (Cao et al., 2020), (Giannakopoulos et al., 2017), (Aumiller et al., 2022)

Czech: (Giannakopoulos et al., 2015),(Straka et al., 2018), (Varab and Schluter, 2021), (Giannakopoulos et al., 2017), (Ladhak et al., 2020a), (Perez-Beltrachini and Lapata, 2022), (Aumiller et al., 2022)

Danish: (Varab and Schluter, 2020), (Varab and Schluter, 2021), (Aumiller et al., 2022)

Dari: (Palen-Michel and Lignos), (Varab and Schluter, 2021)

Dutch: (Nguyen and Daumé III, 2019), (Varab and Schluter, 2021), (Giannakopoulos et al., 2017), (Ladhak et al., 2020a), (Aumiller et al., 2022), (Tikhonov and Malykh, 2022)

English: (Giannakopoulos et al., 2015), (Cheng and Lapata, 2016; Hermann et al., 2015), (Grusky et al., 2018), (Hasan et al., 2021), (Fabbri et al., 2019), (Cachola et al., 2020), (Takeshita et al., 2024b), (Jiang and Dreyer, 2024), (Saxena and Keller, 2024), (Mahbub et al., 2023), (Urlana et al., 2023), (Datta et al., 2023), (Klaus et al., 2022), (Palen-Michel and Lignos), (Cohen et al., 2021), (Koupaee and Wang, 2018), (Huang et al., 2021), (Sharma et al., 2019), (Napoles et al., 2012),

[3] This list includes both Traditional Chinese and Simplified Chinese

(Nguyen and Daumé III, 2019), (Kornilova and Eidelman, 2019), (Hasler et al., 2003), (Lins et al., 2019), (Zhang and Tetreault, 2019), (Wang and Ling, 2016), (Ganesan et al., 2010), (Cohan et al., 2018), (Kim et al., 2018), (Gliwa et al., 2019), (Zhu et al., 2021), (Yasunaga et al., 2019), (Cao et al., 2016), (Wei and Gao, 2018), (Narayan et al., 2018), (Abacha and Demner-Fushman, 2019), (Yadav et al., 2022), (Xu et al., 2021), (Zhang and Wan, 2017), (Ghalandari et al., 2020), (Sotudeh et al., 2021), (Hayashi et al., 2021), (Malykh et al., 2020), (Ta et al., 2023), (Syed et al., 2020), (Liu et al., 2023), (Kraaij et al., 2005), (Völske et al., 2017), (Chen et al., 2021c), (Gupta et al., 2021), (Kryściński et al., 2021), (Ladhak et al., 2020b), (Hayashi et al., 2020), (Antognini and Faltings, 2020), (Wang et al., 2022a), (Feigenblat et al., 2021), (Lu et al., 2020), (Chen et al., 2021b), (Meng et al., 2021), (Khalman et al., 2021), (Roush and Balaji, 2020), (Zhong et al., 2021), (Bhatia et al., 2014), (Ulrich et al., 2008), (Varab and Schluter, 2021), (Li et al., 2018), (Li et al., 2017), (Liang et al., 2022), (Zhu et al., 2018), (Chen and Zhuge, 2018), (Yadav et al., 2021), (Gorinski and Lapata, 2015), (Verma et al., 2023), (Sanabria et al., 2018), (Wang et al., 2021), (Zopf, 2018), (Zopf et al., 2016), (Collins et al., 2017), (Lev et al., 2019), (Park and Park, 2023), (Giannakopoulos et al., 2017), (Ladhak et al., 2020a), (Zhu et al., 2019), (Perez-Beltrachini and Lapata, 2022), (Wang et al., 2022b), (Aumiller et al., 2022), (Bhattacharjee et al., 2021), (Takeshita et al., 2024a), (Fatima and Strube, 2021), (Tikhonov and Malykh, 2022), (Xiaorui, 2023)

Esperanto: (Varab and Schluter, 2021), (Giannakopoulos et al., 2017)

Estonian: (Aumiller et al., 2022)

Filipino: (Varab and Schluter, 2021)

Finnish: (Giannakopoulos et al., 2017), (Aumiller et al., 2022)

French: (Giannakopoulos et al., 2015), (Hasan et al., 2021), (Palen-Michel and Lignos), (Scialom et al., 2020), (Nguyen and Daumé III, 2019), (Varab and Schluter, 2021), (Liang et al., 2022), (Eddine et al., 2020), (Verma et al., 2023), (Wang et al., 2021), (Giannakopoulos et al., 2017), (Ladhak et al., 2020a), (Perez-Beltrachini and Lapata, 2022), (Bhattacharjee et al., 2021), (Tikhonov and Malykh, 2022), (Xiaorui, 2023)

Fulah: (Varab and Schluter, 2021)

Georgian: (Palen-Michel and Lignos), (Varab and Schluter, 2021), (Giannakopoulos et al., 2017)

German: (Scialom et al., 2020), (Nguyen and Daumé III, 2019), (Varab and Schluter, 2021), (Kew et al., 2023), (Wang et al., 2021), (Zopf, 2018), (Giannakopoulos et al., 2017), (Wang et al., 2022b), (Aumiller et al., 2022), (Takeshita et al., 2024a), (Fatima and Strube, 2021), (Tikhonov and Malykh, 2022)

Gujarati: (Hasan et al., 2021), (Satapara et al., 2022), (Urlana et al., 2023), (Varab and Schluter, 2021), (Liang et al., 2022), (Verma et al., 2023), (Ladhak et al., 2020a), (Perez-Beltrachini and Lapata, 2022), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Haitian: (Palen-Michel and Lignos), (Varab and Schluter, 2021)

Hausa: (Hasan et al., 2021), (Palen-Michel and Lignos), (Varab and Schluter, 2021), (Liang et al., 2022), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Hebrew: (Giannakopoulos et al., 2015), (Varab and Schluter, 2021), (Mondshine et al., 2024)

Hindi: (Giannakopoulos et al., 2015), (Hasan et al., 2021), (Satapara et al., 2022), (Urlana et al., 2023), (Datta et al., 2023), (Varab and Schluter, 2021), (Liang et al., 2022), (Verma et al., 2023), (Wang et al., 2021), (Ladhak et al., 2020a), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Hungarian: (Varab and Schluter, 2021), (Barta et al., 2024), (Aumiller et al., 2022)

Icelandic: (Varab and Schluter, 2021)

Igbo: (Hasan et al., 2021), (Varab and Schluter, 2021), (Liang et al., 2022), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Indonesian: (Hasan et al., 2021), (Palen-Michel and Lignos), (Koto et al., 2020), (Kurniawan and Louvan, 2018), (Varab and Schluter, 2021), (Liang et al., 2022), (Verma et al., 2023), (Wang et al., 2021), (Giannakopoulos et al., 2017), (Ladhak et al., 2020a), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Irish: (Varab and Schluter, 2021), (Aumiller et al., 2022)

Italian: (Landro et al., 2022), (Nguyen and Daumé III, 2019), (Varab and Schluter, 2021), (Giannakopoulos et al., 2017), (Ladhak et al., 2020a), (Aumiller et al., 2022), (Takeshita et al., 2024a), (Tikhonov and Malykh, 2022)

Japanese: (Hasan et al., 2021),(Nguyen and Daumé III, 2019), (Varab and Schluter, 2021), (Liang et al., 2022), (Yamamura and Shimada, 2018), (Verma et al., 2023), (Giannakopoulos et al., 2017), (Ladhak et al., 2020a), (Bhattacharjee et al., 2021), (Takeshita et al., 2024a), (Tikhonov and Malykh, 2022), (Xiaorui, 2023)

Javanese: (Giannakopoulos et al., 2017)

Kannada: (Urlana et al., 2023), (Varab and Schluter, 2021)

Kinyarwanda: (Palen-Michel and Lignos), (Varab and Schluter, 2021)

Kirghiz: (Hasan et al., 2021), (Varab and Schluter, 2021), (Liang et al., 2022), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Korean: (Hasan et al., 2021), (Palen-Michel and Lignos), (Liang et al., 2022), (Giannakopoulos et al., 2017), (Ladhak et al., 2020a), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Kurdish: (Palen-Michel and Lignos), (Badawi, 2023), (Varab and Schluter, 2021)

Lao: (Palen-Michel and Lignos), (Varab and Schluter, 2021)

Latvian: (Varab and Schluter, 2021), (Giannakopoulos et al., 2017), (Aumiller et al., 2022)

Limburgish: (Giannakopoulos et al., 2017)

Lingala: (Varab and Schluter, 2021)

Lithuanian: (Varab and Schluter, 2021), (Aumiller et al., 2022)

Macedonian: (Palen-Michel and Lignos), (Nguyen and Daumé III, 2019), (Varab and Schluter, 2021)

Malagasy: (Nguyen and Daumé III, 2019), (Varab and Schluter, 2021)

Malay: (Urlana et al., 2023), (Giannakopoulos et al., 2017)

Malayalam: (Urlana et al., 2023), (Varab and Schluter, 2021)

Maltese: (Aumiller et al., 2022)

Manipuri: (Urlana et al., 2023)

Marathi: (Hasan et al., 2021), (Urlana et al., 2023), (Urlana et al., 2023), (Varab and Schluter, 2021), (Liang et al., 2022), (Verma et al., 2023), (Giannakopoulos et al., 2017), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Modern Greek: (Giannakopoulos et al., 2015), (Palen-Michel and Lignos), (Nguyen and Daumé III, 2019), (Varab and Schluter, 2021), (Giannakopoulos et al., 2017), (Aumiller et al., 2022)

Mongolian: (Varab and Schluter, 2021)

Nepali: (Hasan et al., 2021), (Varab and Schluter, 2021), (Liang et al., 2022), (Verma et al., 2023), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

North Ndebele: (Palen-Michel and Lignos), (Varab and Schluter, 2021)

Norwegian: (Giannakopoulos et al., 2017)

Oriya: (Urlana et al., 2023), (Varab and Schluter, 2021)

Oromo: (Hasan et al., 2021), (Varab and Schluter, 2021), (Liang et al., 2022), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Panjabi: (Hasan et al., 2021), (Urlana et al., 2023), (Varab and Schluter, 2021), (Liang et al., 2022), (Verma et al., 2023), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Pashto: (Hasan et al., 2021), (Palen-Michel and Lignos), (Varab and Schluter, 2021), (Liang et al., 2022), (Verma et al., 2023), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Persian: (Hasan et al., 2021), (Palen-Michel and Lignos), (Liang et al., 2022), (Farahani et al., 2021), (Giannakopoulos et al., 2017), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Polish: (Varab and Schluter, 2021), (Giannakopoulos et al., 2017), (Aumiller et al., 2022), (Tikhonov and Malykh, 2022)

Portuguese: (Hasan et al., 2021), (Palen-Michel and Lignos), (Nguyen and Daumé III, 2019), (Cardoso et al., 2011), (Paiola et al., 2024), (Varab and Schluter, 2021), (Liang et al., 2022), (Verma et al., 2023), (Sanabria et al., 2018), (Wang et al.,

2021), (Giannakopoulos et al., 2017), (Ladhak et al., 2020a), (Aumiller et al., 2022), (Bhattacharjee et al., 2021), (Tikhonov and Malykh, 2022), (Xiaorui, 2023)

Romanian: (Giannakopoulos et al., 2015), (Varab and Schluter, 2021), (Giannakopoulos et al., 2017), (Aumiller et al., 2022)

Rundi: (Hasan et al., 2021), (Varab and Schluter, 2021), (Liang et al., 2022), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Russian: (Hasan et al., 2021), (Palen-Michel and Lignos), (Scialom et al., 2020), (Nguyen and Daumé III, 2019), (Gusev, 2020), (Varab and Schluter, 2021), (Liang et al., 2022), (Verma et al., 2023), (Wang et al., 2021), (Giannakopoulos et al., 2017), (Ladhak et al., 2020a), (Bhattacharjee et al., 2021), (Tikhonov and Malykh, 2022), (Xiaorui, 2023)

Scottish Gaelic: (Hasan et al., 2021), (Varab and Schluter, 2021), (Liang et al., 2022), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Serbian: [4] (Hasan et al., 2021), (Palen-Michel and Lignos), (Varab and Schluter, 2021), (Liang et al., 2022), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Shona: (Palen-Michel and Lignos), (Varab and Schluter, 2021)

Sinhala: (Hasan et al., 2021), (Varab and Schluter, 2021), (Liang et al., 2022), (Verma et al., 2023), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Slovak: (Ondrejova and Šuppa, 2024), (Šuppa and Adamec, 2020), (Varab and Schluter, 2021), (Giannakopoulos et al., 2017), (Aumiller et al., 2022)

Slovenian: (Varab and Schluter, 2021), (Aumiller et al., 2022)

Somali: (Hasan et al., 2021), (Palen-Michel and Lignos), (Varab and Schluter, 2021), (Liang et al., 2022), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Spanish: (Giannakopoulos et al., 2015), (Hasan et al., 2021), (Soriano et al., 2022), (Palen-Michel and Lignos), (Scialom et al., 2020), (Nguyen and Daumé III, 2019), (Varab and Schluter, 2021), (Liang et al., 2022), (Verma et al., 2023), (Wang et al., 2021), (Giannakopoulos et al., 2017), (Ladhak et al., 2020a), (Bhattacharjee et al., 2021), (Tikhonov and Malykh, 2022), (Xiaorui, 2023)

[4]This list includes Cyrillic and Latin writing systems

Swedish: (Varab and Schluter, 2021), (Aumiller et al., 2022), (Tikhonov and Malykh, 2022)

Tagalog: (Giannakopoulos et al., 2017)

Tajik: (Varab and Schluter, 2021)

Tamil: (Hasan et al., 2021), (Urlana et al., 2023), (Varab and Schluter, 2021), (Liang et al., 2022), (Verma et al., 2023), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Telugu: (Hasan et al., 2021), (Urlana et al., 2023), (Urlana et al., 2022), (Varab and Schluter, 2021), (Liang et al., 2022), (Verma et al., 2023), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Tetum: (Varab and Schluter, 2021)

Thai: (Hasan et al., 2021), (Palen-Michel and Lignos), (Varab and Schluter, 2021), (Liang et al., 2022), (Giannakopoulos et al., 2017), (Ladhak et al., 2020a), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Tibetan: (Palen-Michel and Lignos), (Varab and Schluter, 2021)

Tigrinya: (Hasan et al., 2021), (Palen-Michel and Lignos), (Varab and Schluter, 2021), (Liang et al., 2022), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Turkish: (Hasan et al., 2021), (Palen-Michel and Lignos), (Scialom et al., 2020), (Varab and Schluter, 2021), (Liang et al., 2022), (Wang et al., 2021), (Giannakopoulos et al., 2017), (Ladhak et al., 2020a), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Ukrainian: (Hasan et al., 2021), (Palen-Michel and Lignos), (Galeshchuk, 2023), (Varab and Schluter, 2021), (Liang et al., 2022), (Verma et al., 2023), (Wang et al., 2021), (Giannakopoulos et al., 2017), (Bhattacharjee et al., 2021), (Tikhonov and Malykh, 2022), (Xiaorui, 2023)

Urdu: (Hasan et al., 2021), (Urlana et al., 2023), (Palen-Michel and Lignos), (Faheem et al., 2024), (Varab and Schluter, 2021), (Liang et al., 2022), (Verma et al., 2023), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Uzbek: (Hasan et al., 2021), (Palen-Michel and Lignos), (Varab and Schluter, 2021), (Liang et al., 2022), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Vietnamese: (Hasan et al., 2021), (Palen-Michel and Lignos), (Nguyen et al., 2019), (Tran et al., 2020), (Varab and Schluter, 2021), (Liang et al., 2022), (Wang et al., 2021), (Ladhak et al., 2020a), (Bhattacharjee et al., 2021), (Tikhonov and Malykh, 2022), (Xiaorui, 2023)

Welsh: (Hasan et al., 2021), (Varab and Schluter, 2021), (Liang et al., 2022), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

West African Pidgin English: (Hasan et al., 2021), (Liang et al., 2022), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Xhosa: (Varab and Schluter, 2021)

Yoruba: (Hasan et al., 2021), (Varab and Schluter, 2021), (Liang et al., 2022), (Bhattacharjee et al., 2021), (Xiaorui, 2023)

Swahili: (Hasan et al., 2021), (Palen-Michel and Lignos), (Nguyen and Daumé III, 2019), (Varab and Schluter, 2021), (Liang et al., 2022), (Bhattacharjee et al., 2021), (Xiaorui, 2023)