MassiveSumm: a very large-scale, very multilingual, newswire summarisation dataset

Daniel Varab and Natalie Schluter
IT University of Copenhagen
Denmark
{djam,natschluter}@itu.dk

Abstract

Current research in automatic summarisation is unapologetically anglo-centered–a persistent state-of-affairs, which also predates neural net approaches. High-quality automatic summarisation datasets are notoriously expensive to create, posing a challenge for any language. However, with digitalisation, archiving, and social media advertising of newswire articles, recent work has shown how, with careful methodology application, large-scale datasets can now be simply gathered instead of written. In this paper, we present a large-scale multilingual summarisation dataset containing articles in 92 languages, spread across 28.8 million articles, in more than 35 writing scripts. This is both the largest, most inclusive, existing automatic summarisation dataset, as well as one of the largest, most inclusive, ever published datasets for any NLP task. We present the first investigation on the efficacy of resource building from news platforms in the low-resource language setting. Finally, we provide some first insight on how low-resource language settings impact state-of-the-art automatic summarisation system performance.

1 Introduction

Automatic summarisation datasets are generally expensive to create, because they generally involve a human reading a document several times and then crafting a fluent piece of text that captures both the important information of the document and the intention of the resulting summary. Each datapoint in such a dataset could take hours to manually create. With digitalisation, archiving, and social media advertising of newswire articles, recent work has shown how, with dedicated time and methodology application, large-scale datasets can now be simply gathered instead of written (Grusky et al., 2018; Hermann et al., 2015). But the method development was carried out over English, and until the research presented here, the method has only been applied to a very limited number of relatively richly-resourced languages (Varab and Schluter, 2020; Scialom et al., 2020).

We have extended the methodology further (Section 3) and applied it carefully and widely to generate MassiveSumm: a very large-scale, very multilingual summarisation dataset of 28.8 million articles, containing data in 92 languages, using more than 35 writing scripts. This is by far both the largest, most inclusive, existing automatic summarisation dataset, as well as one of the largest, most inclusive, ever published datasets for any NLP task. The bulk of this paper outlines the size, diversity and inclusivity of the dataset as an automatic summarisation dataset, as well as simply raw text data in comparison with two other multilingual large-scale widely used datasets in NLP: Wikipedia and Common Crawl (Section 4).

In light of extending and applying the data acquisition method under the low-resource setting, we identify some unreasonable conditions for language inclusion in automatic summarisation research, which stand to perpetuate a lack of language diversity in system development and therefore unequal access to these tools. We also present some experimental evidence that failure to include a more diverse set of language data in automatic summarisation research can result in only very language specific system design when language agnostic design has been claimed (Section 5).

2 Related Work

A number of works presenting large-scale datasets for automatic summarisation have been presented in the past couple of years. We survey this work here to provide some research context for MassiveSumm.

The New York Times Corpus (**NYT**) consists of 1.8 million articles from the New York Times (Sandhaus, 2008) between 1987 and 2007. The automatic summarisation portion of this dataset

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 10150–10161
November 7–11, 2021.

consists of 650,000 article-summary pairs, where the summaries are written by library scientists. Unlike the rest of the datasets discussed in this section, NYT is created and maintained by the platform that the articles belong to.

The CNN/Daily Mail (**CNNDM**) dataset (Hermann et al., 2015) is an English language automatically acquired Question Answering dataset composed of newswire articles and their corresponding highlights from two separate platforms: `cnn.com` and `dailymail.co.uk`. The dataset was later converted into a summarisation dataset by concatenating these article highlights into article summaries (Cheng and Lapata, 2016; Nallapati et al., 2016). The summarisation dataset consists of 312,000 summary-article pairs. It has become the most broadly used automatically collected English summarisation dataset.

With the same methodology as CNNDM, Narayan et al. (2018) collected the **XSum** dataset of approximately 230,000 summary-article pairs from the `bbc.com` news platform. And Scialom et al. (2020) collected the **MLSum** dataset for five languages from five corresponding news platforms: French, German, Spanish, Russian, Turkish, catering their platform dependent method to each separate news platform. The resulting datset contains a total of around 1.5 million article-summary language pairs. MLSum was the first large-scale multilingual dataset, but all five of the languages of the dataset were still European, Indo-European, and relatively high-resourced within NLP. We note that while, similarly to XSum, MassiveSumm also contains article-summary pairs from the `bbc.com` platform, there are two important differences which make for zero overlap between the two datasets: (1) we include no English datapoints in our dataset, and (2) our summaries are not article highlights, but social media article descriptions, as is done for the remaining newswire datasets surveyed here.

The **Newsroom** dataset (Grusky et al., 2018) is the first large-scale English dataset generated specifically for automatic summarisation. The key insight into automatically creating this dataset was in observing use of a social media standard, called Open Graph[1], by publishers to improve their search engine results. According to this standard, a description of the article contents, used for advertising on social media, should be recorded in the mark-up of the article's web page. The method allowed for scraping news articles from any news outlet, so long as the news outlet upheld the social media standard. Hence, by contrast to the method for acquiring the CNNDM, Newsroom's method was website agnostic, which meant that scraping was no longer constrained to collecting data from specific platforms. Grusky et al. (2018) created Newsroom by conducting a scrape of news articles from 38 English language news outlets spanning two decades starting from the late 1990s, when news platforms first began digitalising their content widely, to 2017. The dataset contains 1.3 million document summary pairs.

Varab and Schluter (2020) extend, streamline and improve the Newsroom methodology to assemble the first automatic summarisation dataset for Danish, **DaNewsroom**. Their work comprises the first non-English website agnostic approach to large-scale article-summary collection, across 19 Danish news platforms and resulting in a dataset of 1.1M article-summary pairs. The methodology of this paper is adapted from this extension of the Newsroom methodology.

Related to this, the **GlobalVoices** dataset (Nguyen and Daumé III, 2019), is an automatic summarisation dataset across 15 languages from one single platform, `https://globalvoices.org`. Although its original collection is similar to Newsroom and DaNewsroom, the resulting dataset is relatively small with less than 30,000 article-summary pairs across all languages in total, including English. Moreover, approximately 800 English summaries are further crowdsourced. The dataset contains purely parallel data and its intended use is for cross-lingual summarisation. MassiveSumm most likely includes all non-English datapoints scraped for GlobalVoices, as this was one of the hundreds of its news platform data sources.

Two further large-scale datasets are not based on newswire. (1) **BigPatent** (Sharma et al., 2019) consists of 1.3 million U.S. patent English language abstract-document pairs, written between 1971 and 2018, across nine technological areas, all from the Google Patents Public Datasets (Google, 2018). (2) **LCSTS** The Large Scale Chinese Short Text Summarization Dataset (Hu et al., 2015) consists of 2.4 million text-summary pairs from the Sina Weibo microblogging platform, where post texts are paired with summaries provided by the author of each text.

[1] `https://ogp.me/`

Contemporaneously to our work, Hasan et al. (2021) developed **XL-Sum**, a summarisation dataset from the BBC news platform. However, their work covers less than a twelfth of the article-summary pairs: around 1 million across 44 languages and a single news platform, compared with our 12.3 million across 92 languages and 370 news platforms.

3 Methodology

Our methodology consists of roughly three parts: (1) manual annotation, (2) automatic collection, and (3) quality control. The first part is unique to the dataset presented here and represents a work-intensive annotation process which seeks to ensure both breadth in terms of language inclusivity, quality and consistency of the data. The remaining parts are measured adjustments of the prior extensions of Grusky et al. (2018)'s methodology by Varab and Schluter (2020).

Manual annotation. We first compiled a list of languages to be represented in the dataset. Our goal was to cover as many languages as possible, with a prioritisation of breadth, linguistic diversity, and language inclusivity, over depth. Then we manually searched for as many news platforms as possible for each language, by contrast to Grusky et al. (2018) who collected news platforms from publicly available lists.

For each news platform we required either (1) that it published exclusively in the language we had associated with it, or (2) published in way such that we could reliably distinguish the difference between languages later on (for example, the platform identified the languages for us). All other platforms were discarded.

Having determined which news platform were suitable language-wise, the next step was to manually investigate which platforms were technically *suitable*: we required these platforms to point to explicit lists of articles on their platform to avoid non-article content such as frontpages, albums or videos. In total, 370 different platforms met our requirements and were retained.

Automatic collection. With the list of suitable news platforms, we obtained all article URLs for each platform by retrieving them from `archive.org`. This is a slow process.

Having had collected the URLs for each platform we observe a significant difference between the amount of URLs across languages, some in the tens of millions, some in the thousands. We stored article URLs of the language together in language bins. We shuffled each bin and proceed to sample an equal amount of URLs from each bin and output them to a download queue. This allowed us to ensure that less frequent languages would always be scraped at the same priority as more frequent ones. Less frequent languages were sampled until they were exhausted, and thus over represented languages were sub-sampled.

Quality Control. We carry out a number of automatic checks for quality control, similarly to Varab and Schluter (2020). The number of articles filtered out of the dataset due to these checks can be seen in Table 1. In particular, we filter out articles with no contents, summaries with no contents, summaries that are prefixes of the article body, and summaries that are prefixes followed by "...". We quantify this filtering process in Section 4.

Distribution. Practically speaking, the publically available dataset is distributed as a list of urls for each language (split into train/dev/test sets) and a single software package for downloading and processing the web pages.[2]

4 The numbers

Total counts. We refer to Table 1. Over 31 million articles were scraped from 370 news platforms, across 92 languages, from 38 language genera withing the following 16 language families: (1) Indo-European, (2) Afro-Asiatic, (3) Mande, (4) Niger-Congo, (5) Austronesian, (6) Altaic, (7) Sino-Tibetan, (8) Austro-Asiatic, (9) Kartvelian, (10) Uralic, (11) Japanese, (12) Dravidian, (13) Korean, (14) Tai-Kadai, (15) other, for Haïtian, and (16) Aymaran.[3] Of these, approximately 3 million scraped article pages had an empty article (2,981,925) and were filtered from the dataset, leaving over 28.8 million articles of raw multilingual text data, which we refer to as **MassiveSumm-All**.

As explained in Section 3, a number of filters were applied to the dataset to improve its quality for automatic summarisation. In particular, we did a check to ensure that summaries were neither empty nor just prefixes of the article, so that the

[2] `https://github.com/danielvarab/massive-summ`

[3] We took language family definitions and genus definitions from `https://wals.info/` database.

	language	genus (family)	empty src	empty tgt	prefix	ellipsis	ellipsis|prefix	all-prefix	all-ellipsis	all	count	%invalid	valid count
						AFRICA							
2	Swahili	Bantoid (F4)	10,219	48,054	52,166	93,395	144,911	151,246	110,383	202,762	302,565	67.01%	99,803
3	Hausa	West Chadic (F13)	22,753	27,319	42,966	34,402	77,355	84,289	93,015	127,242	233,608	54.47%	106,366
4	Somali	Lowland East Cushitic (F13)	18,112	1,385	39,122	121,122	160,235	138,866	57,903	177,979	204,717	86.94%	26,738
5	Afrikaans	Germanic (F11)	374	8	121,056	5,549	126,173	5,927	121,434	126,551	198,792	63.66%	72,241
6	Kinyarwanda	Bantoid (F4)	17,791	6,878	40,893	21,241	62,062	45,307	65,477	86,128	92,674	92.94%	6,546
7	Amharic	Semitic (F13)	12,247	3,945	21,694	2,002	23,483	17,952	37,675	39,433	84,732	46.54%	45,299
8	North Ndebele	Bantoid (F4)	26,731	7	10,267	1,988	12,209	28,660	37,004	38,881	51,202	75.94%	12,321
9	Shona	Bantoid (F4)	25,130	5	12,505	715	13,205	25,840	37,638	38,330	46,681	82.11%	8,351
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
						EURASIA							
20	Russian	Slavic (F11)	26,564	27,482	432,521	91,252	491,426	145,096	486,458	545,270	1,284,433	42.45%	739,163
21	Spanish	Romance (F11)	36,434	101,844	85,805	428,547	513,726	564,728	223,487	649,907	1,216,217	53.44%	566,310
22	Ukrainian	Slavic (F11)	29,968	37,652	358,697	243,248	598,286	302,697	424,432	657,735	1,252,150	52.53%	594,415
23	Persian	Iranian (F11)	16,277	147,711	428,787	44,699	470,156	195,272	579,432	620,729	1,150,653	53.95%	529,924
24	Arabic	Semitic (F13)	44,039	216,084	403,561	6,296	408,247	263,573	661,071	665,524	1,186,870	56.07%	521,346
25	Chinese	Chinese (F9)	838,069	62,003	36,335	388,542	424,829	1,016,062	890,620	1,052,349	1,171,189	89.85%	118,840
26	German	Germanic (F11)	23,358	246,308	323,190	15,901	333,184	284,787	592,185	602,070	1,080,213	55.74%	478,143
27	Urdu	Indic (F11)	19,236	2,291	469,175	4,213	472,516	25,514	490,602	493,817	1,115,555	44.27%	621,738
28	Hindi	Indic (F11)	6,388	1,059	469,614	34,754	502,814	41,977	477,057	510,037	1,073,514	47.51%	563,477
29	French	Romance (F11)	31,711	112,622	249,625	323,869	564,598	458,696	388,211	699,425	1,007,129	69.45%	307,704
30	Polish	Slavic (F11)	6,808	39,910	435,591	22,334	454,093	68,471	482,246	500,230	983,252	50.88%	483,022
31	Vietnamese	Viet-Muong (F3)	532,441	21,410	125,609	81,298	199,344	590,681	672,481	708,727	920,166	77.02%	211,439
32	Bulgarian	Slavic (F11)	22,272	6,606	273,851	9,206	281,857	37,558	302,351	310,209	977,769	31.73%	667,560
33	Tamil	Southern Dravidian (F14)	1,074	11,654	703,881	126,331	829,332	138,242	715,826	841,243	886,482	94.90%	45,239
34	Hungarian	Ugric (F5)	17,332	28,724	220,577	1,229	221,511	43,082	262,478	263,364	885,749	29.73%	622,385
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
						INTERNATIONAL							
86	Esperanto	Constructed (F11)	0	0	27	103	130	103	27	130	565	23.01%	435
						NORTH AMERICA							
87	Haitian	Creoles and Pidgins (F6)	5,890	12	8,346	3,240	11,582	9,118	14,246	17,460	26,009	67.13%	8,549
						PAPUNESIA							
88	Indonesian	Malayo-Sumbawan (F12)	57,358	7,899	131,349	81,850	213,191	146,982	196,586	278,323	492,909	56.47%	214,586
89	Filipino	Greater Central Philippine (F12)	5	0	40	52	92	57	45	97	294	32.99%	197
90	Tetum	Central Malayo-Polynesian (F12)	0	0	2	0	2	0	2	2	15	13.33%	13
91	Bislama	Creoles and Pidgins (F6)	3	0	0	0	0	3	3	3	4	75.00%	1
						SOUTH AMERICA							
92	Aymara	Aymaran (F0)	32	0	110	104	213	129	142	238	827	28.78%	589
		totals	2,981,925	1,775,581	10,315,099	5,145,760	15,238,148	9,404,789	14,891,856	19,497,177	31,940,180	58.04%	12,443,003

LANGUAGE FAMILY LEGEND
Aymaran (F0)
Kartvelian (F1)
Altaic (F2)
Austro-Asiatic (F3)
Niger-Congo (F4)
Uralic (F5)
other (F6)
Japanese (F7)
Tai-Kadai (F8)
Sino-Tibetan (F9)
Mande (F10)
Indo-European (F11)
Austronesian (F12)
Afro-Asiatic (F13)
Dravidian (F14)

Table 1: Excerpt language article-summary pair counts from Table 8 in the Appendix. In the table columns, **empty_art** is the number of articles with no contents, **empty_sum** is the number of summaries with not contents, **prefix** is the number of summaries that also prefixes of the article, **ellipsis** is the number of summaries that are prefixes of the article followed by "...", **ellipsis|prefix** is the number of either ellipsis or prefix summaries (they are not mutually exclusive), **all-prefix** is the number of summaries after filtering, but including prefixes, **all-ellipsis** is the number of summaries after filtering, but including ellipsis, **all** is the number of empty, prefix or ellipsis summaries (they are not mutually exclusive), **count** is the total number of article-summary pairs, **%invalid** is the proportion of filtered article-summary pairs (all/count), and **valid count** is the number of article-summary pairs after filtering.

resulting dataset did not include trivial instances for system development. MassiveSumm can therefore be seen under two views: **MassiveSumm-All (MS-All)** which consists of all non-empty articles (and any available summaries) before application of the above-mentioned filters. And a subset of this– the **MassiveSumm (MS)** summarisation dataset intended for automatic summarisation system development; this dataset is the result of the application of the filters.

We observe (Table 2) that the majority of the dataset, approximately 16.5 million article-summary pairs, did not survive the summary quality control filtering process. The result was 12,368,113 article-summary pairs surviving a minimal quality control for utility in automatic summarisation system development, of which the automatic-summarisation dataset portion of MassiveSumm consists.

dataset	description	size
MassiveSumm (MS)	Fully filtered automatic summarisation data.	12,368,113 article-summary pairs.
MassiveSumm-All (MS-All)	All non-empty articles scraped.	28,879,290 articles.

Table 2: Summary of the contents of MassiveSumm.

This filtering process resulted in a handful of languages having virtually no presence in the automatic summarisation portion of MassiveSumm. For instance, over 98.7% of Xhosa article-summary pairs were filtered out of the summarisation portion of the dataset, leaving only 172 instances.

Table 3 gives an overview of the article/article-summary pair counts. We note that the Indo-European languages provide the majority of the data in the dataset. The Uralic family (here, only with Hungarian) is also relatively heavily represented in the dataset. The 10 Niger-Congo languages as a whole have less data than a single Indo-European language on average. In Section 5 we discuss why our current methodology can only result in perpetuating such under-representation in dataset quantities.

Comparing with web-scrape multilingual datasets. We compared the intersection of our dataset with two large-scale web datasets widely used by the NLP community: Wikipedia[4] and Common Crawl[5]. An overview of this comparison can be found in Table 4. The manual care that we took in curating the list of platforms from which we wanted to collect data resulted in more data from an improved diversity of languages.

For 52 of our languages, MS-All either matches or surpasses the number of Wikipedia pages for the language in question, showing the importance of the full dataset simply as raw data. In fact, the majority of MassiveSumm languages from South Saharan Africa (14/18) have more documents in MS-All than in Wikipedia. And well over half of the MassiveSumm languages for Eurasia (38/63) have more documents in MS-All than in Wikipedia.

Turning to Common Crawl, almost half of the languages from South Saharan Africa (8/18) have more pages in MS-All than in Common Crawl. Six out of 63 Eurasian languages have more articles in MS-All than in Common Crawl.

When we consider even just the heavily filtered automatic summarisation portion of the data, MS, we find that 10 of the South Saharan African languages contain more pages than Wikipedia, and 5 out of 18 of these languages contain more data than Common Crawl. For Eurasia, 19 of the 63 languages contain more pages than Wikipedia.

Table 5 gives the proportions of the articles in MS-All that are also contained in Common Crawl, for those languages where more than 49% can be obtained. This is 18 languages–around a fifth of the languages represented by MassiveSumm. Hence observe that large portions of easily indexible and crawlable, publicly available, diverse linguistic data are not being scraped into one of the most important datasets for NLP, both in size, but in determining to a large extent which languages get mainstream NLP research: Common Crawl.

5 Reflections on Low-Resource Language Automatic Summarisation

The central datasets for automatic summarisation have consistently been for English. In this section we consider how this focus on English has resulted in limited dataset curation methodology development (Section 5.1) and limited automatic summarisation system design (Section 5.2).

[4] https://en.wikipedia.org/wiki/List_of_Wikipedias#Edition_detailsasofMay10,2021

[5] April 2021 crawl CC-MAIN-2021-04 https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.csv

family	MS-All	%(MS-All)	MS	%(MS)	num langs	MS-All ave	MS-All ave%	MS ave	MS ave%
Indo-European	20990245	72.68%	9062565	73.27%	48	437296.77	1.51%	188803.44	1.53%
Dravidian	2005933	6.95%	333765	2.7%	4	501483.25	1.74%	83441.25	0.67%
Afro-Asiatic	1753871	6.07%	816504	6.6%	7	250553.0	0.87%	116643.43	0.94%
Uralic	868417	3.01%	622385	5.03%	1	868417.0	3.01%	622385.0	5.03%
Altaic	835649	2.89%	362191	2.93%	5	167129.8	0.58%	72438.2	0.59%
Austro-Asiatic	477331	1.65%	257197	2.08%	2	238665.5	0.83%	128598.5	1.04%
Niger-Congo	467630	1.62%	142921	1.16%	10	46763.0	0.16%	14292.1	0.12%
Austronesian	462877	1.6%	232510	1.88%	4	115719.25	0.4%	58127.5	0.47%
Sino-Tibetan	434543	1.5%	177373	1.43%	3	144847.67	0.5%	59124.33	0.48%
Tai-Kadai	252073	0.87%	132287	1.07%	2	126036.5	0.44%	66143.5	0.53%
Kartvelian	182743	0.63%	132055	1.07%	1	182743.0	0.63%	132055.0	1.07%
Japanese	125625	0.44%	87220	0.71%	1	125625.0	0.44%	87220.0	0.71%
other	20120	0.07%	8550	0.07%	2	10060.0	0.03%	4275.0	0.03%
Mande	1438	0.0%	1	0.0%	1	1438.0	0.0%	1.0	0.0%
Aymaran	795	0.0%	589	0.0%	1	795.0	0.0%	589.0	0.0%
Totals	28879290		12368113		92				

Table 3: Language family-wise article counts and proportions for MassiveSumm-All (All) and for the MassiveSumm automatic summarisation dataset (MS).

5.1 Impact on dataset curation

The methodology we use for acquiring this dataset is based on Newsroom (Grusky et al., 2018), a dataset for English. In order for the method to be effective at obtaining data, at least the following two assumptions must be met.

Assumption 1. Digitalisation. Digitised newswire text must be publicly available online for the language, and in sufficiently large quantities. This is not the case, however. For example, a broad manual search for online news platforms in Africa[6] revealed relatively few non-colonial language platforms for the region. Digitised newswire is also sparse or non-existent in, for example, non-standard Arabic dialects, European languages such as Irish or Welsh, as well as indigenous languages in North and South America, and Australia. Hence focus on a strategy created for a language where there are massive amounts of online data, and lack of development of new techniques to acquire data for languages that do not have such an online presence will reinforce the lack of representation of these languages in automatic summarisation research.

Assumption 2. Web page structure conventions. Online news platforms must ensure that their article mark-ups abide by the Open Graph protocol (Cf. Section 3). However, extensive manual inspection revealed that while this is the norm for English and in general for languages of rich western countries, this is not the norm in general. For instance, due to this problem we had to exclude a number of other South Saharan African languages including Southern Sotho, Pulaar, Zulu, and Luganda. Further, as we observe in Table 1, approximately 2 million documents are excluded from MS due to their summaries being empty–the news platforms in the corresponding languages have the correct template structure for their web pages, but do not use them as intended.

In order to develop the know-how to achieve true language diversity in datasets for automatic summarisation (and other NLP tasks), methods for acquiring automatic summarisation data should be developed which do not make these two assumptions. The difference in existence and in quantities of data for the languages of MassiveSumm reflect this requirement, which currently favours Indo-European languages.

5.2 Systems: Low-resource baselines

MassiveSumm provides a means to check whether there is evidence of some impact of a focus on English data for neural automatic summarisation.

The languages. We consider a minimal set of non-Indo-European languages to provide such evidence according to three separate considerations: (1) The languages should have large native speaker populations.[7]. (2) The languages should be non-Indo-European. (3) The set of languages should exhibit different complexity in morphology. (4) The datasets should be of significantly different sizes. (5) Finally, all languages must have readily available word segmentisers.

The set of languages we chose for our experiments all have a population far beyond that of the

[6] https://www.w3newspapers.com/africa/

[7] According to `https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers`

language	family	MS	MS-all	Wiki	CC	MS/Wiki	MS/CC	MS-All/Wiki	MS-All/CC
					AFRICA				
Amharic	Afro-Asiatic	45,299	72,485	14,910	95,305	**303.82%**	47.53%	**486.15%**	76.06%
Bambara	Mande	1	1,438	693	0	0.14%	-	**207.50%**	-
Fulah	Niger-Congo	40	499	278	0	14.39%	-	**179.50%**	-
Hausa	Afro-Asiatic	106,366	210,855	6,829	54,355	**1557.56%**	**195.69%**	**3087.64%**	**387.92%**
Igbo	Niger-Congo	4,341	5,085	2,084	9,728	**208.30%**	44.62%	**244.00%**	52.27%
Lingala	Niger-Congo	1,489	4,429	3,184	5,224	46.77%	28.50%	**139.10%**	84.78%
North Ndebele	Niger-Congo	12,321	24,471	0	0	-	-	-	-
Oromo	Afro-Asiatic	5,816	14,926	1,050	15,432	**553.90%**	37.69%	**1421.52%**	96.72%
Rundi	Niger-Congo	3,646	24,085	618	2,882	**589.97%**	**126.51%**	**3897.25%**	**835.70%**
Shona	Niger-Congo	8,351	21,551	6,660	11,559	**125.39%**	72.25%	**323.59%**	**186.44%**
Somali	Afro-Asiatic	26,738	186,605	5,944	159,270	**449.83%**	16.79%	**3139.38%**	**117.16%**
Swahili	Niger-Congo	99,803	292,346	60,725	243,070	**164.35%**	41.06%	**481.43%**	**120.27%**
Tigrinya	Afro-Asiatic	7,978	18,533	208	25,369	**3835.58%**	31.45%	**8910.10%**	73.05%
Xhosa	Niger-Congo	172	12,876	1,182	37,430	14.55%	0.46%	**1089.34%**	34.40%
					EURASIA				
Albanian	Indo-European	156,336	680,535	82,309	1,296,319	**189.94%**	12.06%	**826.81%**	52.50%
Arabic	Afro-Asiatic	521,346	1,142,831	1,102,405	19,101,195	47.29%	2.73%	**103.67%**	5.98%
Armenian	Indo-European	168,453	807,817	281,101	1,050,372	59.93%	16.04%	**287.38%**	76.91%
Azerbaijani	Altaic	140,685	301,134	177,955	1,548,046	79.06%	9.09%	**169.22%**	19.45%
Bengali	Indo-European	124,351	191,712	103,686	2,681,993	**119.93%**	4.64%	**184.90%**	7.15%
Bosnian	Indo-European	45,575	254,737	84,968	1,311,659	53.64%	3.47%	**299.80%**	19.42%
Bulgarian	Indo-European	667,560	955,497	269,103	9,070,911	**248.07%**	7.36%	**355.07%**	10.53%
Central Khmer	Austro-Asiatic	45,758	89,606	8,230	300,772	**555.99%**	15.21%	**1088.77%**	29.79%
Czech	Indo-European	551,443	609,257	473,960	36,586,487	**116.35%**	1.51%	**128.55%**	1.67%
Dari	Indo-European	20,220	59,199	0	0	-	-	-	-
Georgian	Kartvelian	132,055	182,743	148,069	1,269,380	89.18%	10.40%	**123.42%**	14.40%
Gujarati	Indo-European	43,830	450,740	29,481	294,393	**148.67%**	14.89%	**1528.92%**	**153.11%**
Hindi	Indo-European	563,477	1,067,126	145,723	4,185,074	**386.68%**	13.46%	**732.30%**	25.50%
Hungarian	Uralic	622,385	868,417	483,555	18,592,776	**128.71%**	3.35%	**179.59%**	4.67%
Kannada	Dravidian	47,676	281,630	26,789	309,943	**177.97%**	15.38%	**1051.29%**	90.87%
Kurdish	Indo-European	28,008	94,916	37,232	204,372	75.23%	13.70%	**254.93%**	46.44%
Lao	Tai-Kadai	40,316	53,193	3,594	103,238	**1121.76%**	39.05%	**1480.05%**	51.52%
Latvian	Indo-European	7,080	454,915	105,928	2,970,478	6.68%	0.24%	**429.46%**	15.31%
Lithuanian	Indo-European	326,082	884,547	201,003	5,362,226	**162.23%**	6.08%	**440.07%**	16.50%
Macedonian	Indo-European	86,647	219,869	112,077	889,870	77.31%	9.74%	**196.18%**	24.71%
Malayalam	Dravidian	121,568	634,601	71,996	676,894	**168.85%**	17.96%	**881.44%**	93.75%
Marathi	Indo-European	127,838	476,870	69,262	496,649	**184.57%**	25.74%	**688.50%**	96.02%
Modern Greek	Indo-European	95,023	401,315	188,407	18,299,263	50.43%	0.52%	**213.00%**	2.19%
Nepali	Indo-European	23,993	218,138	31,745	805,140	75.58%	2.98%	**687.16%**	27.09%
Oriya	Indo-European	28,582	388,961	15,592	122,957	**183.31%**	23.25%	**2494.62%**	**316.34%**
Panjabi	Indo-European	83,147	322,520	35,218	168,347	**236.09%**	49.39%	**915.78%**	**191.58%**
Persian	Indo-European	529,924	1,134,376	767,776	20,893,043	69.02%	2.54%	**147.75%**	5.43%
Pushto	Indo-European	58,038	215,927	11,807	90,702	**491.56%**	63.99%	**1828.80%**	**238.06%**
Scottish Gaelic	Indo-European	15,012	16,528	15,198	48,315	98.78%	31.07%	**108.75%**	34.21%
Sinhala	Indo-European	12,252	32,851	16,818	215,962	72.85%	5.67%	**195.33%**	15.21%
Slovak	Indo-European	78,639	581,873	235,863	12,240,989	33.34%	0.64%	**246.70%**	4.75%
Tamil	Dravidian	45,239	885,408	134,646	1,444,153	33.60%	3.13%	**657.58%**	61.31%
Telugu	Dravidian	119,282	204,294	70,641	573,248	**168.86%**	20.81%	**289.20%**	35.64%
Thai	Tai-Kadai	91,971	198,880	142,059	11,108,049	64.74%	0.83%	**140.00%**	1.79%
Tibetan	Sino-Tibetan	1,236	6,455	5,949	32,107	20.78%	3.85%	**108.51%**	20.10%
Ukrainian	Indo-European	594,415	1,222,182	1,073,297	12,688,368	55.38%	4.68%	**113.87%**	9.63%
Urdu	Indo-European	621,738	1,096,319	160,631	725,101	**387.06%**	85.75%	**682.51%**	**151.20%**
Welsh	Indo-European	53,802	154,844	132,464	358,792	40.62%	15.00%	**116.90%**	43.16%

Table 4: Languages for which MassiveSumm carries more raw documents than Wikipedia or Common Crawl.

average European country. And yet two of these languages are severely lower resourced in NLP in general, if not zero-resourced. The languages are:

- **Arabic**, a semitic language with a complex morphology and around 310 million native speakers. We used 432,384 article-summary pairs from MS.
- **Telugu**, a Dravidian language with a moderately rich morphology and around 82 million native speakers. We used 12,633 article-summary pairs from MS.
- **Hausa**, an Afro-Asiatic tonal language with a relatively simple morphology and around 43 million native speakers. We used 78,633 article-summary pairs from MS.

The datasets were split into train/test/dev sets with corresponding proportions 80%/10%/10%. For tokenisation of Arabic and Telugu we used Spacy (Honnibal et al., 2020), and the English tokeniser from NLTK (Loper and Bird, 2002) for Hausa. For sentence segmentation we use `pySBD` (Sadvilkar and Neumann, 2020) for Arabic, and NLTK for the remaining Hausa and Telugu.

The system. OpenNMT's (Klein et al., 2017) reimplementation of the Pointer-Generator system (See et al., 2017) provides efficient state-of-the-

language	%	language	%
Tibetan	96.98%	Lingala	72.05%
Lao	95.45%	Malagasy	67.18%
Bambara	95.37%	Tigrinya	66.69%
Dari	94.87%	Bosnian	63.71%
Rundi	84.28%	Scot. Gaelic	63.71%
Burmese	81.51%	Hungarian	61.30%
Haitian	79.50%	Slovenian	58.21%
Oromo	77.93%	Bislama	50.00%
Kurdish	74.77%	Irish	49.27%

Table 5: Languages from MassiveSumm-All for which the percentage of articles that can also be found in Common Crawl is greater than 49%.

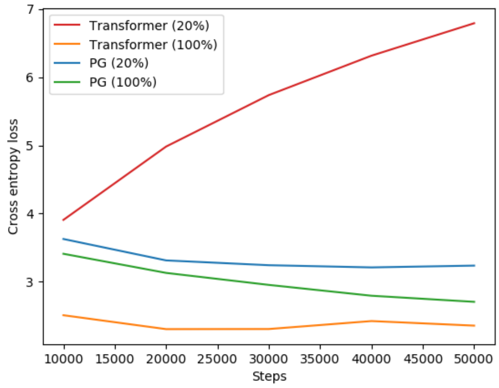

Figure 1: Two fixed architecture configurations run under two data settings: (1) 100% of the training set, and (2) 20% of the training set. The PG model (rnn) is robust to different data settings while the transformer quickly overfits the training data. Loss in the graph is measured over the development set.

art-competitive performance and proved more robust to limits in dataset size than a Transformer (Vaswani et al., 2017) model during our hyperparameter search preparatory experiments–this was a crucial requirement for our low-resource language experiments. We experimented with training both Pointer-Generator and Transformer models over different quantities, 20% and 100% (respectively, 57,444 and 287,227 instances), of CNNDM training data. While the transformer outperforms PG when training on the full dataset (Table 6), it grossly overfits when faced with only 20% of the data for training (Figure 1).

system (train prop.)	R1	R2	RL
Transformer (100% data)	39.06	17.02	36.09
Transformer (20% data)	32.23	11.12	29.99
PG (100%)	38.41	16.31	35.21
PG (80%)	38.18	16.3	35.08
PG (60%)	38.13	16.16	34.92
PG (40%)	38.05	16.13	34.9
PG (20%)	36.81	15.36	33.7

Table 6: Rouge-1, Rouge-2, and Rouge-L (Lin, 2004) scores for comparing Transformer and RNN (PG) models on different proportions of CNNDM training data in preparation for lower-resource language experiments.

For further context, we also train and test on the Newsroom corpus. Since the Newsroom corpus did not filter prefix and ellipsis summaries, we include scores with and without these data filters. We use an 80%/10%/10% split of Newsroom before and after filtering: respectively 994,446/109,147/109,147 and 808,727/88,657/88,768 article-summary pairs.

During training we truncate articles to 400 tokens and summaries to 100 tokens. We fix the random seed but refrain from tying the input and output embeddings (Press and Wolf, 2016). The vocabularies are fixed to 30,000 tokens across all languages and we used no subword tokeniser. At inference time we decoded with a beam size of 10, discarded summaries with less than 35 tokens, block trigrams and apply length penalty with the value $\alpha = 0.9$ (Wu et al., 2016). For further details of the model, we refer to the original papers of (See et al., 2017; Gehrmann et al., 2018) as well as OpenNMT's documentation[8]. Our experiments should act as lower bounds as we conducted no tuning on any of the MassiveSumm datasets.

We include the Lead-3 baseline which simply copies the first three sentences from the article. It is a notoriously strong baseline for automatic summarisation systems and acts as a baseline point of reference that is resilient to training set size limitations.

The results are given in Table 7. In particular, we notice that ROUGE scores tend to be rather low for the largest non-English dataset, Arabic, with the most complex morphology, despite being the largest of the three. As expected, Telugu with the smallest dataset, also has low ROUGE scores. On the other hand, Lead-3 performs better but similarly low in ROUGE score. On the other hand, ROUGE scores for Hausa are significantly higher in scale than Newsroom scores and also significantly outperform the strong Lead-3 baseline. We have 3

[8] https://opennmt.net/OpenNMT-py/examples/Summarization.html

different linguistic contexts and three quite different behaviours, which provides clear evidence that robust development in automatic summarisation must adjust and consider linguistic diversity.

dataset (system)	R1	R2	RL
Arabic (Point.-Gen.)	13.58	4.02	13.53
Arabic (Lead-3)	11.34	3.18	11.27
Hausa (Point.-Gen.)	38.55	28.5	31.91
Hausa (Lead-3)	30.55	17.95	26.68
Telugu (Point.-Gen.)	5.62	1.43	5.62
Telugu (Lead-3)	8.87	2.17	8.7
Newsroom (Point.-Gen.)	34.73	21.25	30.39
Newsroom (Lead-3)	31.12	21.4	28.49
Filt. Newsroom (Point.-Gen.)	28.95	15.5	23.9
Filt. Newsroom (Lead-3)	25.49	14.17	22.49

Table 7: Baseline ROUGE scores for Arabic, Hausa, and Telugu. ROUGE scores for Newsroom added for context.

6 Concluding Remarks

In this paper, we presented the most large-scale, most language and linguistically diverse and inclusive dataset for automatic summarisation to date: MassiveSumm. In acquiring MassiveSumm, we also acquired one of the most diverse and inclusive sources of raw linguistic data to date. We also provided evidence how focus on anglo-centric data acquisition method development and system development were detrimental to both language inclusion and language agnostic system behaviour.

References

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Google. 2018. Google patents public datasets: connecting public, paid, and private patent data.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

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, Lisbon, Portugal. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, 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, Berlin, Germany. Association for Computational Linguistics.

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. In *Proceedings of the 2018*

Conference on Empirical Methods in Natural Language Processing, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Khanh Nguyen and Hal Daumé III. 2019. Global Voices: Crossing borders in automatic news summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 90–97, Hong Kong, China. Association for Computational Linguistics.

Ofir Press and Lior Wolf. 2016. Using the output embedding to improve language models. *arXiv preprint arXiv:1608.05859*.

Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic sentence boundary disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.

Evan Sandhaus. 2008. The new york times annotated corpus. Technical report, Linguistic Data Consortium, Philadelphia.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. 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, Vancouver, Canada. Association for Computational Linguistics.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Daniel Varab and Natalie Schluter. 2020. DaNewsroom: A large-scale Danish summarisation dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6731–6739, Marseille, France. European Language Resources Association.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

A Appendix

This appendix contains the full version of Table 1.

	language	genus (family)	empty src	empty tgt	prefix	ellipsis	ellipsis|prefix	all-prefix	all-ellipsis	all	count	%invalid	valid count
	AFRICA												
1	Swahili	Bantoid (F4)	10,219	48,054	52,166	93,395	144,911	151,246	110,383	202,762	302,565	67.01%	99,803
2	Hausa	West Chadic (F13)	22,753	27,319	42,966	34,402	77,355	84,289	93,015	127,242	233,608	54.47%	106,366
3	Somali	Lowland East Cushitic (F13)	18,112	1,385	39,122	121,122	160,235	138,866	57,903	177,979	204,717	86.94%	26,738
4	Afrikaans	Germanic (F11)	374	8	121,056	5,549	126,173	5,927	121,434	126,551	198,792	63.66%	72,241
5	Kinyarwanda	Bantoid (F4)	17,791	6,878	40,893	21,241	62,062	45,307	65,477	86,128	92,674	92.94%	6,546
6	Amharic	Semitic (F13)	12,247	3,945	21,694	2,002	23,483	17,952	37,675	39,433	84,732	46.54%	45,299
7	North Ndebele	Bantoid (F4)	26,731	7	10,267	1,988	12,209	28,660	37,004	38,881	51,202	75.94%	12,321
8	Shona	Bantoid (F4)	25,130	5	12,505	715	13,205	25,840	37,638	38,330	46,681	82.11%	8,351
9	Rundi	Bantoid (F4)	7,478	74	8,576	11,840	20,416	19,341	16,126	27,917	31,563	88.45%	3,646
10	Tigrinya	Semitic (F13)	6,625	77	4,261	6,262	10,522	12,920	10,957	17,180	25,158	68.29%	7,978
11	Oromo	Lowland East Cushitic (F13)	7,315	14	1,811	7,325	9,135	14,615	9,139	16,425	22,241	73.85%	5,816
12	Malagasy	Barito (F12)	23	0	3,741	5,603	9,318	5,616	3,764	9,331	27,045	34.50%	17,714
13	Xhosa	Bantoid (F4)	124	2	236	12,483	12,718	12,593	360	12,828	13,000	98.68%	172
14	Bambara	Western Mande (F10)	6,137	1	1,437	14	1,451	6,137	7,574	7,574	7,575	99.99%	1
15	Yoruba	Defoid (F4)	33	10	639	555	1,193	588	676	1,226	7,438	16.48%	6,212
16	Igbo	Igboid (F4)	8	492	171	83	253	582	670	752	5,093	14.77%	4,341
17	Lingala	Bantoid (F4)	305	0	2,938	5	2,943	307	3,243	3,245	4,734	68.55%	1,489
18	Fulah	Peul-Serer (F4)	0	215	38	207	244	422	253	459	499	91.98%	40
	EURASIA												
19	Russian	Slavic (F11)	26,564	27,482	432,521	91,252	491,426	145,096	486,458	545,270	1,284,433	42.45%	739,163
20	Spanish	Romance (F11)	36,434	101,844	85,805	428,547	513,726	564,728	223,487	649,907	1,216,217	53.44%	566,310
21	Ukrainian	Slavic (F11)	29,968	37,652	358,697	243,248	598,286	302,697	424,432	657,735	1,252,150	52.53%	594,415
22	Persian	Iranian (F11)	16,277	147,711	428,787	44,699	470,156	195,272	579,432	620,729	1,150,653	53.95%	529,924
23	Arabic	Semitic (F13)	44,039	216,084	403,561	6,296	408,247	263,573	661,071	665,524	1,186,870	56.07%	521,346
24	Chinese	Chinese (F9)	838,069	62,003	36,335	388,542	424,829	1,016,062	890,620	1,052,349	1,171,189	89.85%	118,840
25	German	Germanic (F11)	23,358	246,308	323,190	15,901	333,184	284,787	592,185	602,070	1,080,213	55.74%	478,143
26	Urdu	Indic (F11)	19,236	2,291	469,175	4,213	472,516	25,514	490,602	493,817	1,115,555	44.27%	621,738
27	Hindi	Indic (F11)	6,388	1,059	469,614	34,754	502,814	41,977	477,057	510,037	1,073,514	47.51%	563,477
28	French	Romance (F11)	31,711	112,622	249,625	323,869	564,598	458,696	388,211	699,425	1,007,129	69.45%	307,704
29	Polish	Slavic (F11)	6,808	39,910	435,591	22,334	454,093	68,471	482,246	500,230	983,252	50.88%	483,022
30	Vietnamese	Viet-Muong (F3)	532,441	21,410	125,609	81,298	199,344	590,681	672,481	708,727	920,166	77.02%	211,439
31	Bulgarian	Slavic (F11)	22,272	6,606	273,851	9,206	281,857	37,558	302,351	310,209	977,769	31.73%	667,560
32	Tamil	Southern Dravidian (F14)	1,074	11,654	703,881	126,331	829,332	138,242	715,826	841,243	886,482	94.90%	45,239
33	Hungarian	Ugric (F5)	17,332	28,724	220,577	1,229	221,511	43,082	262,478	263,364	885,749	29.73%	622,385
34	Lithuanian	Baltic (F11)	335	100,060	131,465	327,472	458,586	427,686	231,826	558,800	884,882	63.15%	326,082
35	Armenian	Armenian (F11)	15,906	15,450	107,732	531,897	639,295	547,872	124,117	655,270	823,723	79.55%	168,453
36	Kannada	Southern Dravidian (F14)	502,488	5,767	205,491	44,631	246,047	535,026	711,609	736,442	784,118	93.92%	47,676
37	Italian	Romance (F11)	3,172	227,502	20,405	26,676	46,996	256,974	250,819	277,294	885,915	31.30%	608,621
38	Albanian	Albanian (F11)	4,524	9,133	509,787	6,381	515,192	19,912	523,416	528,723	685,059	77.18%	156,336
39	Malayalam	Southern Dravidian (F14)	4,125	169	118,459	395,556	513,530	399,184	122,743	517,158	638,726	80.97%	121,568
40	Czech	Slavic (F11)	148	1,010	52,020	5,143	56,826	6,279	53,156	57,962	609,405	9.51%	551,443
41	Slovak	Slavic (F11)	668	25,599	477,616	118,930	477,946	144,886	503,576	503,902	582,541	86.50%	78,639
42	Marathi	Indic (F11)	925	7,158	332,233	9,749	341,935	17,771	340,308	349,957	477,795	73.24%	127,838
43	Gujarati	Indic (F11)	422	2,035	6,455	398,661	405,103	400,890	8,882	407,332	451,162	90.29%	43,830
44	Oriya	Indic (F11)	37,874	36,075	177,446	180,032	357,429	220,856	219,146	398,253	426,835	93.30%	28,582
45	Modern Greek	Greek (F11)	4,755	10,094	232,570	69,587	296,865	83,769	246,787	311,047	406,070	76.60%	95,023
46	Turkish	Turkic (F2)	5,172	2,453	254,896	5,308	257,354	11,759	261,427	263,805	376,891	70.00%	113,086
47	Portuguese	Romance (F11)	21,362	36,428	72,366	107,697	179,571	165,336	130,131	237,210	374,602	63.32%	137,392
48	Latvian	Baltic (F11)	7,950	10,535	444,534	2,746	444,752	13,779	455,567	455,785	462,865	98.47%	7,080

language family
Aymaran (F0)
Kartvelian (F1)
Altaic (F2)
Austro-Asiatic (F3)
Niger-Congo (F4)
Uralic (F5)
other (F6)
Japanese (F7)
Tai-Kadai (F8)
Sino-Tibetan (F9)
Mande (F10)
Indo-European (F11)
Austronesian (F12)
Afro-Asiatic (F13)
Dravidian (F14)

Table 8: Full table of language article-summary pair counts. In the table columns, **empty_art** is the number of articles with no contents, **empty_sum** is the number of summaries with not contents, **prefix** is the number of summaries that also prefixes of the article, **ellipsis** is the number of summaries that are prefixes of the article followed by "...", **ellipsis|prefix** is the number of either ellipsis or prefix summaries (they are not mutually exclusive), **all-prefix** is the number of summaries after filtering, but including prefixes, **all-ellipsis** is the number of summaries after filtering, but including ellipsis, **all** is the number of empty, prefix or ellipsis summaries (they are not mutually exclusive), **count** is the total number of article-summary pairs, **%invalid** is the proportion of filtered article-summary pairs (all/count), and **valid count** is the number of article-summary pairs after filtering. (Table continued on next page.)

	language	genus (family)	empty src	empty tgt	prefix	ellipsis	ellipsis|prefix	all-prefix	all-ellipsis	all	count	%invalid	valid count
		EURASIA CONT'D											
49	Azerbaijani	Turkic (F2)	3,910	737	143,471	16,894	159,757	21,496	148,099	164,359	305,044	53.88%	140,685
50	Bosnian	Slavic (F11)	7,284	15,627	92,486	107,925	200,120	124,251	108,842	216,446	262,021	82.61%	45,575
51	Pushto	Iranian (F11)	45,882	27,965	106,804	48,889	155,642	97,018	154,958	203,771	261,809	77.83%	58,038
52	Thai	Kam-Tai (F8)	23,309	13,098	41,198	59,015	96,445	92,788	75,715	130,218	222,189	58.61%	91,971
53	Nepali	Indic (F11)	725	23,181	58,859	112,622	171,476	136,016	82,670	194,870	218,863	89.04%	23,993
54	Macedonian	Slavic (F11)	449	368	87,230	45,901	133,010	46,562	87,947	133,671	220,318	60.67%	86,647
55	Panjabi	Indic (F11)	6,353	2,471	218,108	19,491	237,043	28,174	226,916	245,726	328,873	74.72%	83,147
56	Icelandic	Germanic (F11)	2,167	15	95,095	63,965	158,994	66,081	97,276	161,110	199,970	80.57%	38,860
57	Bengali	Indic (F11)	32,106	940	50,535	16,175	66,577	49,065	83,544	99,467	223,818	44.44%	124,351
58	Japanese	Japanese (F7)	74,711	139	38,179	8,515	46,694	74,937	112,893	113,116	200,336	56.46%	87,220
59	Telugu	South-Central Dravidian (F14)	5,421	2,845	74,573	11,179	83,273	18,339	82,030	90,433	209,715	43.12%	119,282
60	English	Germanic (F11)	184	65	161,571	1,463	162,062	1,696	161,808	162,295	209,021	77.65%	46,726
61	Georgian	Kartvelian (F1)	4,097	2,785	32,106	16,185	48,175	22,795	38,765	54,785	186,840	29.32%	132,055
62	Slovenian	Slavic (F11)	2,269	17	20,113	2,702	22,421	4,977	22,397	24,696	168,688	14.64%	143,992
63	Burmese	Burmese-Lolo (F9)	65,254	190	35,508	7,579	43,045	67,459	100,803	102,925	160,222	64.24%	57,297
64	Welsh	Celtic (F11)	1,915	33	100,617	2,197	101,036	4,118	102,565	102,957	156,759	65.68%	53,802
65	Tajik	Iranian (F11)	501	13	86,260	2,323	88,518	2,837	86,774	89,032	150,419	59.19%	61,387
66	Kurdish	Iranian (F11)	52,706	652	58,405	7,908	66,292	61,230	111,747	119,614	147,622	81.03%	28,008
67	Serbian	Slavic (F11)	6,554	16,395	15,784	38,628	54,248	60,395	38,693	76,015	189,144	40.19%	113,129
68	Uzbek	Turkic (F2)	8,100	376	41,983	3,858	45,544	11,870	50,216	53,556	138,748	38.60%	85,192
69	Hebrew	Semitic (F13)	6	4,011	41	623	664	4,640	4,058	4,681	107,642	4.35%	102,961
70	Central Khmer	Khmer (F3)	8,404	672	15,919	28,557	44,449	36,360	24,338	52,252	98,010	53.31%	45,758
71	Lao	Kam-Tai (F8)	20,719	1,718	10,770	610	11,370	22,836	33,059	33,596	73,912	45.45%	40,316
72	Dari	Iranian (F11)	8,221	463	34,394	4,561	38,950	12,811	43,069	47,200	67,420	70.01%	20,220
73	Croatian	Slavic (F11)	4,525	6,859	804	971	1,746	12,258	12,173	13,033	79,634	16.37%	66,601
74	Assamese	Indic (F11)	257	707	531	37,514	38,044	38,228	1,274	38,758	48,917	79.23%	10,159
75	Tibetan	Bodic (F9)	35,994	10	4,717	1,300	6,017	36,496	40,717	41,213	42,449	97.09%	1,236
76	Romanian	Romance (F11)	1,263	3,822	2,082	1,075	3,143	6,127	7,140	8,195	82,190	9.97%	73,995
77	Sinhala	Indic (F11)	62	67	6,476	14,072	20,541	14,192	6,604	20,661	32,913	62.77%	12,252
78	Kirghiz	Turkic (F2)	31	2	8,331	110	8,421	141	8,362	8,452	31,536	26.80%	23,084
79	Scottish Gaelic	Celtic (F11)	158	4	1,512	0	1,512	162	1,674	1,674	16,686	10.03%	15,012
80	Dutch	Germanic (F11)	8	0	186	143	326	151	194	334	1,805	18.50%	1,471
81	Irish	Celtic (F11)	0	0	1,263	48	1,280	48	1,263	1,280	1,780	71.91%	500
82	Catalan	Romance (F11)	12	0	42	102	143	109	54	150	816	18.38%	666
83	Mongolian	Mongolic (F2)	4	10	429	63	490	76	443	503	647	77.74%	144
84	Swedish	Germanic (F11)	0	0	43	4	47	4	43	47	364	12.91%	317
85	Danish	Germanic (F11)	6	0	52	13	65	19	58	71	337	21.07%	266
		INTERNATIONAL											
86	Esperanto	Constructed (F11)	0	0	27	103	130	103	27	130	565	23.01%	435
		NORTH AMERICA											
87	Haitian	Creoles and Pidgins (F6)	5,890	12	8,346	3,240	11,582	9,118	14,246	17,460	26,009	67.13%	8,549
		PAPUNESIA											
88	Indonesian	Malayo-Sumbawan (F12)	57,358	7,899	131,349	81,850	213,191	146,982	196,586	278,323	492,909	56.47%	214,586
89	Filipino	Greater Central Philippine (F12)	5	0	40	52	92	57	45	97	294	32.99%	197
90	Tetum	Central Malayo-Polynesian (F12)	0	0	2	0	2	0	2	2	15	13.33%	13
91	Bislama	Creoles and Pidgins (F6)	3	0	0	0	0	3	3	3	4	75.00%	1
		SOUTH AMERICA											
92	Aymara	Aymaran (F0)	32	0	110	104	213	129	142	238	827	28.78%	589
		totals	2,981,925	1,775,581	10,315,099	5,145,760	15,238,148	9,404,789	14,891,856	19,497,177	31,940,180	58.04%	12,443,003