

Descartes: Generating Short Descriptions of Wikipedia Articles

Marija Šakota
EPFL
Switzerland
marija.sakota@epfl.ch

Maxime Peyrard
EPFL
Switzerland
maxime.peyrard@epfl.ch

Robert West
EPFL
Switzerland
robert.west@epfl.ch

ABSTRACT

Wikipedia is one of the richest knowledge sources on the Web today. In order to facilitate navigating, searching, and maintaining its content, Wikipedia's guidelines state that all articles should be annotated with a so-called short description indicating the article's topic (e.g., the short description of BEER is “*Alcoholic drink made from fermented cereal grains*”). Nonetheless, a large fraction of articles (ranging from 10.2% in Dutch to 99.7% in Kazakh) have no short description yet, with detrimental effects for millions of Wikipedia users. Motivated by this problem, we introduce the novel task of automatically generating short descriptions for Wikipedia articles and propose Descartes, a multilingual model for tackling it. Descartes integrates three sources of information to generate an article description in a target language: the text of the article in all its language versions, the already-existing descriptions (if any) of the article in other languages, and semantic type information obtained from a knowledge graph. We evaluate a Descartes model trained for handling 25 languages simultaneously, showing that it beats baselines (including a strong translation-based baseline) and performs on par with monolingual models tailored for specific languages. A human evaluation on three languages further shows that the quality of Descartes's descriptions is largely indistinguishable from that of human-written descriptions; e.g., 91.3% of our English descriptions (vs. 92.1% of human-written descriptions) pass the bar for inclusion in Wikipedia, suggesting that Descartes is ready for production, with the potential to support human editors in filling a major gap in today's Wikipedia across languages.

ACM Reference Format:

Marija Šakota, Maxime Peyrard, and Robert West. 2023. Descartes: Generating Short Descriptions of Wikipedia Articles. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 4, 2023, Austin, TX, USA*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3543507.3583220>

1 INTRODUCTION

With over 42M articles in around 300 languages, Wikipedia is the largest encyclopedia ever built. Since most people are unfamiliar with most entities covered by Wikipedia articles, Wikipedia's guidelines stipulate that each article should be annotated by a so-called

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WWW '23, April 30–May 4, 2023, Austin, TX, USA

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-9416-1/23/04...\$15.00
<https://doi.org/10.1145/3543507.3583220>

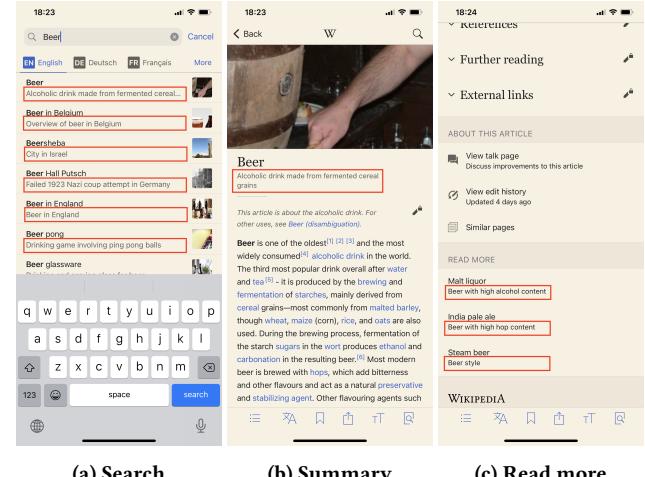


Figure 1: Three use cases of short descriptions on Wikipedia.

short description providing “a very brief indication of the field covered by the article”.¹ For instance, the short description of BEER is “*Alcoholic drink made from fermented cereal grains*”.

When shown together with article titles, short descriptions can vastly facilitate navigation and search, as shown in Fig. 1, which exemplifies how short descriptions are used in the Wikipedia mobile app to (a) disambiguate search results, (b) summarize an article's topic at the top of the page, and (c) annotate links in the “read more” section at the bottom of the page. Beyond knowledge consumption, short descriptions are also useful for knowledge production and management; e.g., editors rely heavily on lists of article titles for organizing their work, and annotating these titles with short descriptions can increase intuition and efficiency.

Although Wikipedia's guidelines require a short description for each article, a large fraction of all articles do not have one, as shown in Table 1 for 25 language editions. The problem is particularly striking for low-resource languages (e.g., 99.7% of Kazakh and 95.1% of Lithuanian articles have no short description), but it also strongly affects high-resource languages (e.g., 19.7% of English and 19.1% of German articles have no short description). Across the 25 languages of Table 1, 9.3M articles (43%) have no short description, with detrimental effects on the navigability, retrievability, and maintainability of Wikipedia's content for hundreds of millions of users.

As volunteer time is notoriously scarce on Wikipedia—there is vastly more work that should than that could be done by human editors [42]—writing the millions of missing short descriptions is not feasible with human labor alone, but requires automated tools. Building these tools is the focus of this paper.

¹ See https://en.wikipedia.org/wiki/Wikipedia:Short_description; regarding the relationship with Wikidata's item descriptions, see Appendix A.

The task of generating short descriptions for Wikipedia articles occupies a sweet spot, being (1) *within reach* yet (2) *challenging*, with the potential for (3) *high-impact* applications.

First, the task is *within reach* as the information required to solve it is usually present somewhere in the respective article.

Second, the task remains *challenging* as standard methods do not suffice: Hand-coded rules (e.g., regular expressions [15]) are insufficient given the extreme heterogeneity of the mapping from article texts to short descriptions; in many cases the ideal description does not appear explicitly as a substring of the article text at all, but needs to be combined from multiple sentences (cf. example of Fig. 1b); and the concept in question may not yet have an article in the target language to begin with from which a short description could be extracted by a rule. Machine translation, which could be used to translate already-existing descriptions of the concept in question from other languages into the target language, is also insufficient, since it is not clear from which of the potentially many languages to translate; ideally, one would like to pool information from the already-existing descriptions across all languages, combined with additional textual modalities such as article texts and non-textual modalities such as knowledge graphs. For these reasons, the task also goes beyond conventional text summarization (cf. Sec. 2).

Finally, tackling the task has the potential for *high impact* as short descriptions are such an important, yet overwhelmingly incomplete, feature of Wikipedia, while the threshold for incorporating auto-generated descriptions in live Wikipedia is low, compared to other kinds of auto-generated content such as entire article texts [4, 10, 23, 35], infoboxes [19, 37, 43], etc.: Wikipedia requires all edits to be vetted by humans, which is feasible with relatively little training (and on small, intuitive user interfaces) for short descriptions (cf. footnote 1), whereas it requires experienced editors (and more complex user interfaces) for other types of content such as entire articles or infoboxes. Incorporating auto-generated descriptions is thus well suited for helping recruit and onboard new editors, instead of further straining the already-overloaded existing editors.

Proposed solution.² We tackle the task with *Descartes* (short for “*Describer of articles*”), a generative language model that integrates three modalities to produce short descriptions of a Wikipedia article about entity E in language L : (1) the texts of all articles about E in all languages (L as well as others), (2) the already-existing descriptions of E in all languages other than L , and (3) information about E 's semantic type obtained from the Wikidata knowledge graph [41].

Results. An automated evaluation on 25 (high- as well as low-resource) languages shows that Descartes outperforms baselines by a wide margin, including a strong baseline built on state-of-the-art machine translation. A single multilingual model rivals the performance of a collection of monolingual models custom-trained for individual languages (implying crosslingual transfer capabilities), and incorporating already-existing descriptions from other languages and knowledge graph information further increases performance. Given the inherent limitations of automated evaluation metrics for language generation tasks such as ours, we also conducted a crowdsourcing-based evaluation with human raters, who

Table 1: Statistics of the 25 Wikipedia language editions considered here. Description length counts tokens for all languages except those marked by *, where it counts characters.

Language	Articles	Missing desc.	Missing desc. (%)	Avg. desc. length
en	English	5204K	1023K	19.65
de	German	2041K	389K	19.07
nl	Dutch	1886K	192K	10.18
es	Spanish	1463K	690K	47.21
it	Italian	1287K	465K	36.14
ru	Russian	1406K	960K	68.25
fr	French	979K	298K	30.45
zh	Chinese	1025K	876K	85.46
ar	Arabic	986K	315K	31.97
vi	Vietnamese	122K	1172K	95.85
ja	Japanese	1103K	858K	77.78
fi	Finnish	451K	300K	66.66
ko	Korean	422K	376K	89.11
tr	Turkish	321K	253K	79.04
ro	Romanian	282K	162K	57.48
cs	Czech	178K	85K	47.89
et	Estonian	195K	160K	81.80
lt	Lithuanian	185K	176K	95.11
kk	Kazakh	220K	219K	99.67
lv	Latvian	92K	71K	77.82
hi	Hindi	130K	80K	61.19
ne	Nepali	29K	25K	85.64
my	Burmese	44K	38K	87.45
si	Sinhala	17K	16K	94.05
gu	Gujarati	29K	7K	25.59

compared Descartes's descriptions to human-generated gold descriptions of the respective articles. Across three languages (English, Hindi, Romanian), the quality of our descriptions is indistinguishable from, or comes close to, that of the gold descriptions. Manual error analysis of the English case further shows that the errors made by Descartes follow a similar distribution as those made by human editors, and even when our description was not preferred by human raters, it still constitutes a high-quality description that could be added to Wikipedia in 60% of cases (a similar rate as *vice versa*, i.e., for human descriptions not preferred by human raters). Overall, 91.3% of Descartes's English descriptions (vs. 92.1% of gold descriptions) meet Wikipedia's quality criteria, which suggests that our model is ready for production.

Contributions. In a nutshell, our contributions are the following.

- (1) We introduce the novel task of short-description generation for Wikipedia articles.
- (2) We propose Descartes, a model that tackles the task by merging multilingual textual information and language-independent semantic type information (Sec. 3).
- (3) In extensive evaluations we demonstrate that the quality of Descartes's descriptions is largely indistinguishable from human-crafted gold descriptions (Sec. 4 and 5).

2 RELATED WORK

Wikipedia summarization. While we are, to the best of our knowledge, the first to tackle the task of short-description generation for Wikipedia articles, researchers have previously attempted to automatically generate various other components of Wikipedia articles, including infoboxes [19, 37, 43], section-based article structures [34], section titles [13], keyphrases [24], article summaries

² Code, data, and models available at <https://github.com/epfl-dlab/descartes>

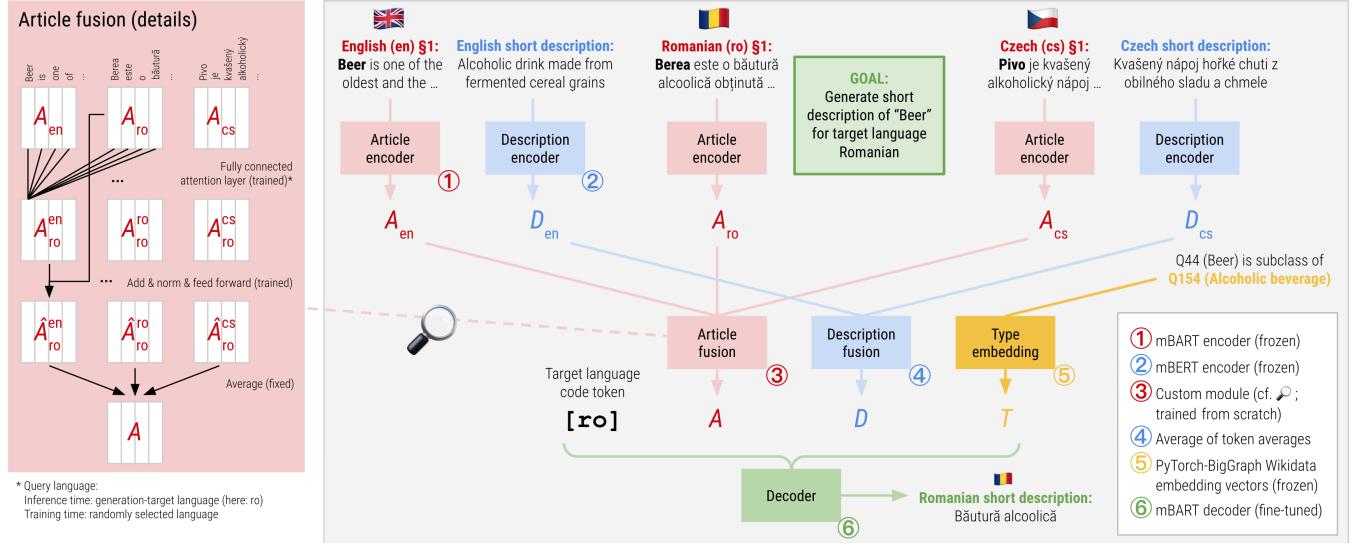


Figure 2: Overview of Descartes, our system for generating short descriptions for Wikipedia articles, as described in Sec. 3, illustrated on the task of generating a Romanian description of the article about BERE (BEER). “§1” stands for “first paragraph”.

[8, 20, 39, 40], or entire articles [4, 10, 23, 35]. Unlike ours, most prior work has focused on English Wikipedia, with notable exceptions relevant to our work including multilingual article summarization methods for low-resource languages [17, 18].

Extreme text summarization. Our work is also related to the area of extreme text summarization, where researchers have aimed at generating one-sentence summaries of news [22, 30, 44] or scientific [6] articles, or function names for code snippets [1]. Although these research goals are similar to ours, we emphasize that, by simultaneously leveraging multiple input modalities and languages (cf. Sec. 1), our method goes beyond conventional summarization methods. Previous extensions to the standard text summarization paradigm include multi-document summarization [11, 27, 28, 36]; methods that enrich input texts with named-entity labels, part-of-speech tags [29], or background information sourced from knowledge graphs [11, 12, 16]; and methods that enforce factual consistency in summaries [7, 14, 46]. In designing Descartes, we were partly inspired by the design principles behind such methods, but go beyond by tailoring our model to the particularities of the task of short-description generation for Wikipedia.

3 METHOD

We begin with a high-level sketch of our method (see overview diagram of Fig. 2) and give further details in the following paragraphs. As mentioned, Descartes integrates three input modalities in order to generate short descriptions for an entity E in language L :

- (1) **Article texts** about E in all languages: Articles in different languages may contain different cues, some more relevant than others for generating short descriptions, and pooling information across languages allows for learning to dynamically select the most relevant information. Note that the mapping from entities to articles is available via Wikidata.

- (2) **Existing descriptions** of E in all languages other than L : Intuitively, translating from other languages should yield good descriptions, and examples in several languages provide even more cues. As different Wikipedia language editions have different norms about short descriptions, the model should learn how to best leverage each language.
- (3) **Semantic type information:** Short descriptions typically capture semantic types (e.g., BEER has type ALCOHOLIC BEVERAGE in Wikidata), so types and descriptions are expected to contain rich cues about one another.

To process these cues, Descartes starts by transforming raw article texts into distributed (matrix-shaped) representations A_l for each language l in which E has an article. The language-specific representations A_l are then fused into a language-independent article representation A . Analogously, raw description texts are transformed into distributed representations D_l , which are then fused into a language-independent description representation D . The semantic type of E is represented via an embedding vector T obtained from the Wikidata knowledge graph. The description (a token sequence) in target language L is then generated by a decoder that receives as input the concatenation of A , D , T , and a special target language token. This way, Descartes can flexibly generate descriptions for any entity E in any target language L , whether L already has an article for E or not. The model is trained to maximize the likelihood of E 's ground-truth description in language L .

The next paragraphs provide more details about the above steps.

Integrating article texts across languages. Distributed language-specific article representations A_l are obtained by feeding raw article texts to the pretrained encoder of the multilingual BART (mBART) language model [25]. Here we consider only the first paragraph of each article, where all the required information tends to be contained. Language-specific article representations A_l are

then fused into a language-independent representation A as follows (cf. red box labeled “Article fusion (details)” in Fig. 2): First, we pick a *query language* q (see below regarding choice of q) and use a transformer-style attention mechanism [38] to fuse A_q in a pairwise fashion with each language l ’s article representation A_l , thus obtaining a pairwise representation A_q^l for every language l . Following the standard transformer attention paradigm [38], we then add the original A_q to every A_q^l (a so-called skip connection), followed by layer normalization and a feed-forward layer, obtaining \hat{A}_q^l . Finally, we average \hat{A}_q^l over all languages l in order to obtain the language-independent article representation A . (For a more formal description, cf. Appendix B.)

During training, we use a randomly sampled language as the query language q , which leads to better generalization than when always using the target language L as q ; e.g., it enables using target languages at inference time that do not even have an article about entity E yet. This said, we found that, during inference, using the target language L as q gives the best results when L has an article about E , so we choose $q = L$ in such cases.

Integrating existing descriptions across languages. Distributed language-specific representations D_l of already-existing descriptions are obtained by feeding raw description texts to the pretrained encoder of the multilingual BERT (mBERT) language model³ and then averaging all token embeddings into a single vector. In order to obtain the language-independent description representation D , we then average D_l over all languages l . We opted for simple averages over tokens and languages because this was found to be as effective as more sophisticated aggregation mechanisms during development. Similarly, we chose mBERT instead of the above-introduced mBART because it is a smaller model but performed well during development.

Integrating semantic types. Type information for Wikipedia articles is readily available from the Wikidata knowledge graph [41], which links each language-specific article (e.g., BEER) to a language-independent entity (e.g., Q44) and specifies semantic types (e.g., Q154, i.e., ALCOHOLIC BEVERAGE) for almost all entities via the INSTANCE OF and SUBCLASS OF relations. To obtain distributed type representations that can be plugged into our neural network model, we use PyTorch-BigGraph’s⁴ [21] Wikidata graph embedding (which maps every Wikidata entity to a 200-dimensional vector) and use as E ’s type representation T the average embedding vector of all Wikidata entities E' for which the relation (E , INSTANCE OF, E') or (E , SUBCLASS OF, E') holds (or, if there is no such E' , a dummy representation equal to the average over all type embedding vectors).

Implementation. We implemented the above model (available as per footnote 2) in PyTorch on top of Hugging Face. The mBART model⁵ supports 25 languages and has 610M parameters (12-layer encoder, 12-layer decoder, 16 heads, 1024-dimensional hidden layers). The mBERT model (cf. footnote 3) supports 102 languages and has 110M parameters (12-layer encoder, 12 heads, 768-dimensional hidden layers). As mBART’s 25 languages form a subset of mBERT’s 102 languages, our implementation of Descartes supports those 25

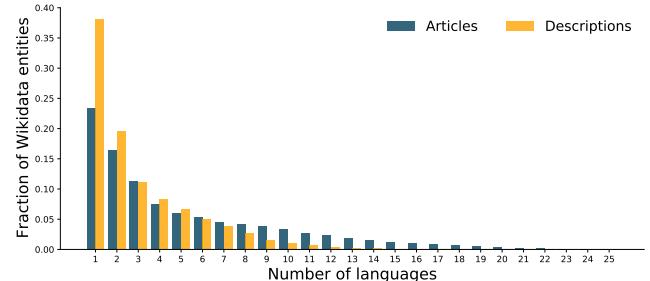


Figure 3: Distribution of number of languages in which Wikidata entities have articles (blue) or descriptions (orange).

languages (cf. Table 1), which span many language families and the full spectrum from low to high resources.

The pretrained encoders of mBART and mBERT, as well as the PyTorch-BigGraph entity embeddings, are frozen and used without fine-tuning, whereas the mBART decoder is finetuned during training and the article fusion module (red box in Fig. 2) is trained from scratch. This way, training remains efficient, with only 215M trainable parameters, compared to 720M for full mBART and mBERT combined. Finetuning on 100K descriptions took 24 hours on a single GPU, and generating all 9.3M descriptions currently missing in the 25 supported languages would take around 40 days on a single GPU (but note that the process can be trivially parallelized and needs to be run only once).

4 AUTOMATIC EVALUATION

To quantify Descartes’s performance, we start with a large-scale evaluation based on automatically computed scores measuring the similarity between auto-generated and human-generated descriptions. Here the latter serve as a gold standard, and perfect performance corresponds to reproducing the human-generated ground truth exactly. This evaluation is imperfect since (1) automatically computed scores may not reflect actual output quality perfectly and (2) it cannot correctly score cases where auto-generated descriptions are better than the human-generated ones, which are taken as perfect by definition. We therefore later complement our analysis with a human evaluation where Descartes’s descriptions are explicitly compared against human-generated ones (Sec. 5).

4.1 Experimental setup

Data. The automatic evaluation was conducted for the 25 languages of Table 1, based on three sets of Wikidata entities sampled uniformly at random under the constraint that all sampled entities have an article and a short description in at least one of the 25 languages: a training set of 100K entities, a testing set of 10K entities, and a validation set (used during development) of 10K entities. During training, each entity from the training set contributed one data point, with a randomly sampled language (out of those for which a description was available) serving as the target language. During testing, we generated descriptions for all languages in which the respective entity had a human-generated ground-truth description.

Fig. 3 shows the distribution of the number of languages in which Wikidata entities from the training set have articles and descriptions.

³ <https://huggingface.co/bert-base-multilingual-uncased>

⁴ <https://github.com/facebookresearch/PyTorch-BigGraph>

⁵ <https://huggingface.co/facebook/mbart-large-cc25>

Table 2: Automatic evaluation in 25 languages (cf. Table 1) via average MoverScores. First row shows percentage of Wikidata entities with article in language. Notes (cf. Sec. 4.1): Translation baseline evaluated only on a subset of the test set (articles with a description in at least one other language; 61.9% of test instances). Summarization baseline (evaluated only for English): .537.

	en	de	nl	es	it	ru	fr	zh	ar	vi	ja	fi	ko	tr	ro	cs	et	lt	kk	lv	hi	ne	my	si	gu
% entities w/ art.	77	45	41	41	37	36	35	27	27	24	24	17	16	14	12	9	7	7	7	5	4	1	1	.7	.7
Baselines																									
Prefix	.564	.563	.572	.565	.558	.622	.556	.627	.651	.581	.594	.584	.618	.581	.560	.567	.596	.616	.622	.570	.636	.645	.710	.750	.785
Translation	.682	.661	.619	.686	.647	.704	.688	.693	.716	.704	.679	.661	.703	.658	.648	.659	.775	.693	.665	.674	.733	.688	.713	.903	.872
Proposed models																									
Descartes	.781	.798	.846	.838	.855	.821	.825	.815	.883	.818	.840	.814	.802	.704	.900	.807	.888	.882	.633	.747	.888	.800	.835	.862	.900
[no desc]	.778	.793	.846	.834	.853	.824	.816	.812	.879	.825	.841	.814	.810	.696	.898	.787	.886	.874	.604	.773	.890	.788	.795	.821	.895
[no types]	.775	.794	.844	.840	.852	.819	.814	.815	.884	.831	.842	.816	.803	.709	.898	.794	.898	.872	.653	.756	.891	.780	.836	.901	.899
[no desc/types]	.772	.776	.840	.836	.847	.824	.810	.809	.881	.837	.833	.812	.800	.697	.894	.794	.881	.892	.614	.753	.888	.771	.772	.923	.892
[monolingual]	.770	.798	.833	.835	.857	.827	.810	.820	.881	.827	.850	.816	.794	.708	.907	.793	.887	.889	.591	.800	.889	.791	.810	.935	.978

Although entities with an article in only one language are most common, 77% have articles in at least two languages. Similarly, although entities with a description in only one language are most common, 62% have descriptions in at least two languages (one of which serves as a target language during training). Additionally, for 94% of entities, a semantic type could be extracted from Wikidata (cf. Sec. 3). These statistics show that all three modalities considered by Descartes (multilingual article texts, multilingual descriptions, semantic types) can be exploited in the majority of cases.

For a sample of descriptions by Descartes, see Appendix D.

Models. In our analysis we compare eight methods: full Descartes as described in Sec. 3, four ablated versions that ignore one or more input modalities, and three baselines:

- (1) **Descartes:** Our full model (cf. Sec. 3).
- (2) **Descartes [no desc]:** Ignoring existing descriptions.
- (3) **Descartes [no types]:** Ignoring semantic type information.
- (4) **Descartes [no desc/types]:** Using only (multilingual) article texts.
- (5) **Descartes [monolingual]:** Using only the article text in the target language. Note that this requires training 25 models, one per target language.
- (6) **Prefix baseline:** Sanity-check baseline that returns the first n words of the article (in the target language) for which a description is to be generated, where n is the average number of words in descriptions in the target language (cf. right-most column of Table 1; for Chinese, Japanese, and Korean, characters were used as a unit instead of words).
- (7) **Translation baseline:** When descriptions are available in other languages, we translate one to the target language using Google Translate. If more than one are available, we choose the highest-resource language (in terms of number of Wikipedia articles). If no descriptions in other languages are available, this baseline cannot be applied.
- (8) **Summarization baseline:** BART [22] finetuned on the XSum extreme summarization dataset [30] (English only).

Performance metric. As the main performance metric for automatic evaluation we use MoverScore [45], which was designed for measuring the semantic similarity between auto-generated text and a ground-truth reference. The MoverScore lies in $[0, 1]$, and

larger values are better. It leverages multilingual contextual representations and is based on the earth mover's distance in embedding space between the two texts. By working in semantic-embedding space rather than surface-token space, MoverScore correlates better with human judgment than token-matching metrics such as BLEU or ROUGE, as was shown on a variety of text generation tasks such as summarization or translation [45]. This fact is particularly important in abstractive settings such as ours, where many good outputs consisting of entirely different words are possible.

4.2 Results

Table 2 shows the test-set performance (average MoverScore) of all eight evaluated methods. We see that Descartes beats the baselines in each language, whereas among the Descartes versions, there is no clear winner across languages. Full Descartes performs best in 6 languages, Descartes [no types] in 7, Descartes [no desc] in 3, Descartes [no desc/types] in 2, and the respective monolingual version optimized for the respective target language in 10 languages. (The sum is greater than 25 due to ties.) The fact that the multilingual Descartes versions perform similarly to, or even better than, the monolingual versions across languages is particularly encouraging, as it implies that there is little to no performance decline in the multilingual setting and that, consequently, we can cater to both high- and low-resource languages with one unified model.

Given that translating an existing description from another language *a priori* appears to be a powerful heuristic, it may be surprising that Descartes outperforms the translation baseline by such large margins (e.g., by 15%/21%/37% in English/German/Dutch, the largest languages in our data). However, different languages often have different conventions regarding description styles, and machine-translating a text consisting of only a few words might be an under-constrained problem. Both issues are alleviated by Descartes: as a supervised model, it can learn language-specific conventions, and article texts as well as multiple existing descriptions can provide disambiguating contexts that further constrain the problem.

Whereas the above column-wise comparisons in Table 2 (of methods for a given language) are valid, row-wise comparisons and averages (of languages for a given method) are more problematic, as MoverScores may not be comparable across languages for various reasons. For instance, scores in low-resource languages may be

Table 3: Pairwise automatic model comparison. Cell (i, j) has probability that model of row i beats model of column j on random test instance, as estimated via Bradley–Terry model (cf. Sec. 4.2). Asterisks (*) mark probabilities statistically significantly ($p < 0.05$) different from 0.5 under sign test.

	Prefix	Descartes	[no desc/types]	[no desc]	[no types]	[monolingual]
Prefix		0.075*	0.077*	0.076*	0.076*	0.076*
Descartes		0.925*	0.522*	0.508*	0.501*	0.495*
[no desc/types]		0.923*	0.478*	0.485*	0.479*	0.490
[no desc]		0.925*	0.492*	0.515*	0.493*	0.492*
[no types]		0.924*	0.499*	0.521*	0.507*	0.496*
[monolingual]		0.924*	0.505*	0.510	0.508*	0.504*

inflated because embeddings of texts in these languages tend to be closer to one another in the multilingual embedding space, which may bias their mutual distances to smaller values. Additionally, Wikipedia articles in low-resource languages cover a narrower range of topics, which makes the description-generation task inherently easier for these languages and may inflate MoverScores.

Pairwise comparison. To circumvent these issues, we deploy a more robust, paired aggregation methodology termed *Pairformance* [33], which uses the Bradley–Terry model [5] to infer a latent performance score $s(m)$ for each model m such that the probability that model m_1 performs better than model m_2 on a randomly drawn test instance (across all languages) is approximated by $\frac{s(m_1)}{s(m_1)+s(m_2)}$. This aggregation is scale-independent, which is important in our case due to the non-comparability of absolute MoverScores across languages. Table 3, which contains a comparison of all models as captured by the above pairwise probabilities,⁶ demonstrates that leveraging existing descriptions and semantic types increases model performance statistically significantly, compared to a model that uses only article texts (Descartes [no desc/types]); e.g., full Descartes beats Descartes [no desc/types] on 52.2% of test instances, while Descartes [no desc] and Descartes [no types] do so on 51.5% and 52.1%, respectively. These results also confirm that, across languages, the unified multilingual Descartes model performs nearly indistinguishably from a collection of language-specific monolingual models (the latter having a win ratio of 50.5%).

The fact that different versions of Descartes perform similarly is partly due to the fact that, as we show in Sec. 5, the model has little room for improvement, given that it is already indistinguishable from human performance. Moreover, the fact that adding existing descriptions to the model gives a similar boost as adding semantic types may be caused by these two cues carrying similar information. By manually inspecting model outputs, we observed, however, that each of these signals can improve the generated descriptions in certain individual cases (examples in Appendix D).

Propensity-score-based analysis. Descartes’s use case is to generate descriptions for articles that do not have one yet. For evaluation, we, however, needed to use articles that already have a description, which may differ systematically from articles that do not have a description yet (e.g., they may be about more popular topics, better written, etc.). If this is the case, our test set would not

⁶The translation baseline is omitted from the table as it cannot handle test instances without descriptions in other languages (cf. Sec. 4.1).

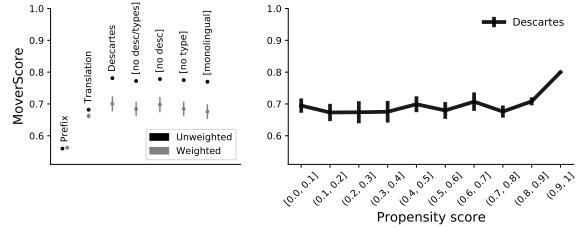


Figure 4: Propensity-score-based analysis analysis of automatic evaluation on English test set. *Left:* Unweighted (black, cf. Table 2) and propensity-score-weighted (gray, for approximating distribution of articles without descriptions) average MoverScores. *Right:* MoverScore of Descartes stratified by propensity score. Error bars: 95% confidence intervals (black error bars in left plot are too small to be visible).

be fully representative of the articles to which our model will be deployed, and the above-reported performance would be a biased estimate.

To correct for this potential mismatch, we employ *propensity scores* [3], a tool commonly used in observational studies in order to correct for such biases. In our case, an article i ’s propensity score p_i is defined as its probability of already having a short description, judging by its textual content only. Propensity scores allow us to reweight [26] the test set, which by construction contains only articles that already have a description, in order to reflect the distribution of articles that do not have a description yet. In particular, by defining article i ’s weight as $(1 - p_i)/p_i$ and taking a weighted average of MoverScores over the test articles allows us to estimate the expected performance for articles without descriptions, even though we cannot directly evaluate on such articles.

Focusing on English, we created a dataset containing randomly sampled English articles, of which around 20% have no description (cf. Table 1), and finetuned a BERT-based [9] binary classifier for predicting if an article i has a description, based on i ’s text. The classifier’s softmax output then serves as i ’s propensity score p_i .

Fig. 4 (left) summarizes the performance of all methods via weighted as well as unweighted average MoverScores (the latter stemming from Table 2). Although all scores are lower in the weighted case, implying that test articles that resemble articles without descriptions (which are upweighted in this analysis) tend to have lower MoverScores, the ordering of methods remains the same as in the unweighted analysis. Fig. 4 (right), which plots the average MoverScore of full Descartes stratified by propensity score, further shows that test performance is stable across the full propensity-score range, with only those test articles with the highest propensity scores standing out with higher-than-average MoverScores. This implies that Descartes works well across the full range of articles.

Limits of automatic evaluation. In interpreting the results of the above automatic evaluation, an important caveat must be kept in mind: low-propensity-score articles can be expected to also have lower-quality human-generated ground-truth descriptions, but the evaluation metric (MoverScore) measures the similarity between auto-generated and ground-truth descriptions. Hence, low Mover-Scores may be due to low-quality model outputs or due to low-quality ground truth. Moreover, even if the ground truth were

perfect, the automatic metric might not reflect similarity to it perfectly, and automatic metrics are only useful for comparing relative improvements between models, rather than measuring whether the outputs are “good enough” to solve the task. In order to overcome these fundamental limitations of any automatic evaluation, we conducted an evaluation with human raters, described next.

5 HUMAN EVALUATION

To surpass the above limitations of automatic metrics, we conducted a human evaluation on Amazon Mechanical Turk (MTurk). Whereas in the automatic evaluation the human-generated description was a gold standard, here it is a competitor, so the auto-generated description is now able to score higher than the human-generated one, which was impossible in the automatic evaluation. We evaluated Descartes (in its full version) on three languages: English, Hindi, and Romanian. These languages were chosen because they span different language families and have different amounts of text available for training, while all being spoken by a large number of MTurk workers [31], which facilitated recruiting.

5.1 Experimental setup

Task design. In each rating task, workers were shown an entire Wikipedia page alongside the auto- and the human-generated descriptions (in random order to avoid positioning bias), and were asked to choose the more appropriate one, in a forced-choice manner, following standard methodological recommendations [32]. The instructions stated that “a good description indicates in one phrase what the article is about” and listed examples of good descriptions borrowed from Wikipedia’s guidelines (cf. footnote 1). Each MTurk task consisted of a batch of 10 binary choices (each for a different Wikipedia article). One of these was a fabricated honeypot, where the auto-generated description was replaced with the ground-truth description of a randomly sampled different article, such that it was obvious which of the two descriptions was better. This allowed us to filter out unreliable workers *post hoc* (see below). Each rating task was done by three different workers, and the winning description per pair was determined by majority vote.

Data. To avoid comparisons between two identical descriptions, we removed from the test set all the articles for which Descartes reproduced the ground truth (up to capitalization), which eliminated 37.5%, 64.6%, 69.0% of articles in English, Hindi, and Romanian, respectively (itself a manifestation of the quality of Descartes’s output). From the rest, we sampled a total of 1000 articles per language, stratifying by MoverScore and sampling 100 articles per decile, with the intent to cover the full performance range according to the automatic evaluation and to thus be able to determine whether the automatic evaluation scores correlate with human choice. For English, sampling within each MoverScore decile was uniformly at random. As Romanian and Hindi Wikipedia have less topical diversity, we drew biased samples with the goal of covering the topical spectrum more evenly, by representing articles via their Wikidata graph embedding (cf. footnote 4) and sampling far-apart data points using the k-means++ [2] cluster seeding algorithm.

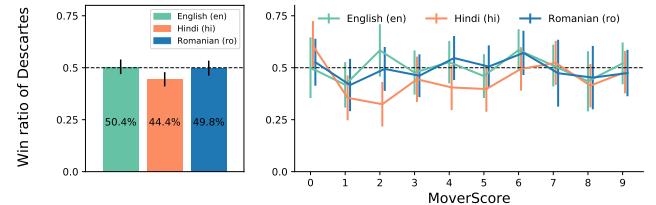


Figure 5: Human evaluation results, showing that Descartes’s descriptions are largely indistinguishable from human-written ground truth. *Left:* Fraction of articles for which human raters preferred Descartes over ground truth. *Right:* *Idem*, stratified by automatically computed similarity (MoverScore) of Descartes and ground truth. Error bars: 95% CIs.

Since, to enforce conciseness, we limited Descartes’s output length when generating descriptions, some (in particular for disambiguation pages) ended abruptly, before the decoder could terminate its output. Such test instances were omitted from the analysis.

Crowdworkers. To ensure reliable ratings, we recruited only workers with at least a minimum number/fraction of previously approved tasks (1000/99% for English; 500/97% for Hindi and Romanian). Hindi and Romanian participants were restricted by location (India and Romania, respectively). We targeted a pay rate of \$8–10 per hour, guided by US minimum wage. In order to verify that workers understood Hindi or Romanian, respectively, they needed to correctly answer a simple multiple-choice entry question in that language. Finally, we filtered unreliable workers by excluding those who failed on over 20% of the encountered honeypots (see above).

5.2 Results

Performance analysis. The results, reported in Fig. 5 (left), show that our descriptions were rated as better than their human-generated counterparts in 50.5% [47.3%, 53.3%], 44.4% [41.6%, 47.3%], and 49.8% [46.7%, 52.9%] of the tested articles in English, Hindi, and Romanian, respectively (95% CIs in brackets; Fleiss’ κ = 0.23, 0.32, 0.78). First, this shows that Descartes performs similarly well for all three languages, despite them having widely different amounts of training data. Second, as implied by the 95% CIs containing 50%, Descartes’s descriptions in English and Romanian are of indistinguishable quality from the human-generated ones, whereas for Hindi the latter were only slightly preferred by raters. Fig. 5 (right), which disaggregates the results by MoverScore decile, shows that performance is high throughout. The lack of correlation between MoverScores and human preference *post hoc* supports the aforementioned doubts regarding the appropriateness of automatic scores such as MoverScore: while the similarity with the ground truth is lower in lower MoverScore bins by definition, the ground truth itself seems to be of lower quality there as well. As a concrete example, in the lowest MoverScore decile, AIN O SALISH KENDRO (a Bangladeshi NGO) has ground truth “Organization”, whereas Descartes generated the better “Human rights organization in Bangladesh”.

Error analysis. To further investigate the quality of auto-generated descriptions and their human-generated counterparts, we manually inspected the descriptions (both auto- and human-generated) that

Table 4: Error categories for human evaluation.

Error category	Description
Too vague	The description does not contain enough information to identify the entity
Factual error	The description contains wrong facts, for example a wrong year, nationality of a person
Formatting error	The description is not an appropriate short-description independent of its content: it contains grammatical errors, formatting problems like using “...”, errors with white spaces
Mis-focused	The description is describing other entities
Too long	The description contains too many details, is too long, or made of several sentences

were not selected by MTurk raters and categorized them based on an error taxonomy developed via iterative coding (see Appendix C for details), shown in Table 4. Note that, due to the forced-choice setting, even a non-preferred description may be—and indeed often is—of high quality, in which case it was labeled as “good enough”.

Using this taxonomy, two authors annotated the non-preferred description of 300 random test articles, without knowing how it was generated ($\text{Fleiss' } \kappa = 0.77$; disagreements were manually resolved). The error distributions for auto- and human-generated descriptions, plotted in Fig. 6 (left), are nearly identical, the main differences being that Descartes makes more factual errors, but never generates overly long descriptions (as output length was limited during decoding). Manual inspection further revealed that factual errors resulted from (1) generating incorrect years or (2) not using the word “former” for outdated properties. (The second type of error was also commonly found in human-generated descriptions.)

We thus obtain the global error profile of Fig. 6 (right), which represents the non-preferred descriptions of Fig. 6 (left) as well as the preferred descriptions and cases where auto- and human-generated descriptions are identical. (The rarest error types are grouped as “other errors” here.) Summing up the three leftmost bars, we conclude that 91.3% of Descartes’s descriptions are either identical to the ground truth, preferred over the ground truth, or non-preferred but still of high quality (“good enough”). For the human ground truth, the analogous fraction is 92.1%. Taken together, this analysis demonstrates that the quality of Descartes’s descriptions is on par with that of human-written descriptions.

Propensity-score-based analysis. Finally, we analyze the results using propensity scores (cf. Sec. 4), to investigate possible biases due to the evaluation being done on articles that already have a description in the target language, whereas the model is intended for articles that do not have one yet. As before (cf. Fig. 4), we stratify the results by propensity score and compute propensity-score-weighted averages (Fig. 7), finding that the fraction of times Descartes’s description was preferred does not vary systematically with the propensity to have a description and is always close to 50%—further evidence that the model works well on its intended use case: to write descriptions for articles that do not have one yet.

6 CONCLUSION

This work addresses a large and consequential gap in Wikipedia: that millions of articles across languages have no short description yet, which hampers the searchability, navigability, and maintainability of the world’s largest encyclopedia for millions of users.

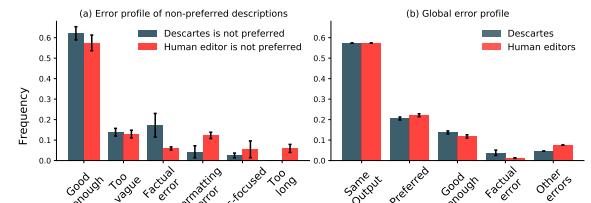


Figure 6: Error analysis on English test set, for (left) descriptions not preferred by human raters and (right) all descriptions, i.e., including preferred descriptions and descriptions that were produced identically by Descartes and human Wikipedia editors. 91.3% of Descartes’s descriptions are either identical to the ground truth, preferred over the ground truth, or non-preferred but still of high quality (“good enough”), similar to the analogous number for ground-truth descriptions (92.1%). Error bars: 95% CIs.

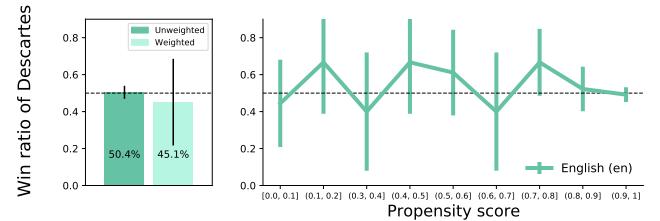


Figure 7: Propensity-score-based analysis of human evaluation on English test set. Left: Unweighted (dark) and propensity-score-weighted (light) average win ratio of Descartes (i.e., fraction of articles where human raters preferred Descartes over ground truth). Right: Win ratio stratified by propensity score. Error bars: 95% CIs.

Our proposed model—Descartes—is, to the best of our knowledge, the first solution to this problem in the literature. Our automatic evaluation demonstrates that a single multilingual model performs better than a strong machine translation baseline and as good as monolingual models that were specifically optimized for individual languages. Our human evaluation went further by showing that Descartes is essentially indistinguishable from human-written descriptions across three languages (English, Hindi, Romanian).

Encouraged by these results, we are currently working on turning Descartes into a practical tool with the goal of supporting human editors in decreasing the otherwise ever-growing number of missing descriptions. We envision a micro-task that can be accomplished even by less experienced editors (perfectly suited as an onboarding exercise for newcomers), e.g., via an interface showing an article snippet alongside Descartes’s top descriptions, from which the editor chooses the best (with the option to reject Descartes’s recommendations and instead contribute another one).

More broadly, this work highlights the potential of solutions built upon state-of-the-art generative language models to help close content gaps in Wikipedia that would otherwise keep widening forever, given the scarcity of volunteer editor time. We are looking forward to seeing more applications in the same spirit.

Acknowledgments. With support from Swiss National Science Foundation (200021_185043), H2020 (952215), Microsoft, Google, and Facebook.

REFERENCES

- [1] Miltiadis Allamanis, Hao Peng, and Charles Sutton. 2016. A Convolutional Attention Network for Extreme Summarization of Source Code. In *Proceedings of The 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 48)*, Maria Florina Balcan and Kilian Q. Weinberger (Eds.). PMLR, New York, New York, USA, 2091–2100. <https://proceedings.mlr.press/v48/allamanis16.html>
- [2] David Arthur and Sergei Vassilvitskii. 2006. *k-means++: The Advantages of Careful Seeding*. Technical Report 2006-13. Stanford InfoLab. <http://ilpubs.stanford.edu:8090/778/>
- [3] Peter C Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* 46, 3 (2011), 399–424.
- [4] Siddhartha Banerjee and Prasenjit Mitra. 2015. WikiCreator: Improving Wikipedia Stubs Automatically. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China, 867–877. <https://doi.org/10.3115/v1/P15-1084>
- [5] Ralph Allan Bradley and Milton E. Terry. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* 39, 3/4 (1952), 324–345. <http://www.jstor.org/stable/2334029>
- [6] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S. Weld. 2020. TLDR: Extreme Summarization of Scientific Documents. arXiv:2004.15011 [cs.CL]
- [7] Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2017. Faithful to the Original: Fact Aware Neural Abstractive Summarization. arXiv:1711.04434 [cs.IR]
- [8] Andrew Chisholm, Will Radford, and Ben Hachey. 2017. Learning to generate one-sentence biographies from Wikidata. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain, 633–642. <https://aclanthology.org/E17-1060>
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [10] Angela Fan and Claire Gardent. 2022. Generating Biographies on Wikipedia: The Impact of Gender Bias on the Retrieval-Based Generation of Women Biographies. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 8561–8576. <https://doi.org/10.18653/v1/2022.acl-long.586>
- [11] Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2019. Using Local Knowledge Graph Construction to Scale Seq2Seq Models to Multi-Document Inputs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 4186–4196. <https://doi.org/10.18653/v1/D19-1428>
- [12] Patrick Fernandes, Miltiadis Allamanis, and Marc Brockschmidt. 2021. Structured Neural Summarization. arXiv:1811.01824 [cs.LG]
- [13] Anjalie Field, Sascha Rothe, Simon Baumgartner, Cong Yu, and Abe Ittycheriah. 2020. A Generative Approach to Titling and Clustering Wikipedia Sections. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*. Association for Computational Linguistics, Online, 79–87. <https://doi.org/10.18653/v1/2020.ngt-1.9>
- [14] Beliz Gunel, Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2020. Mind The Facts: Knowledge-Boosted Coherent Abstractive Text Summarization. arXiv:2006.15435 [cs.CL]
- [15] Marti A. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*. <https://aclanthology.org/C92-2082>
- [16] Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge Graph-Augmented Abstractive Summarization with Semantic-Driven Cloze Reward. arXiv:2005.01159 [cs.CL]
- [17] Lucie-Aimée Kaffee, Hady Elsahar, Pavlos Vougiouklis, Christophe Gravier, Frédérique Laforest, Jonathon Hare, and Elena Simperl. 2018. Learning to Generate Wikipedia Summaries for Underserved Languages from Wikidata. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana, 640–645. <https://doi.org/10.18653/v1/N18-2101>
- [18] Lucie-Aimée Kaffee, Pavlos Vougiouklis, and Elena Simperl. 2021. Using natural language generation to bootstrap missing Wikipedia articles: A human-centric perspective. *Semantic Web* (February 2021). <https://eprints.soton.ac.uk/449718/>
- [19] Dustin Lange, Christoph Böhm, and Felix Naumann. 2010. Extracting Structured Information from Wikipedia Articles to Populate Infoboxes. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (Toronto, ON, Canada) (CIKM '10)*. Association for Computing Machinery, New York, NY, USA, 1661–1664. <https://doi.org/10.1145/1871437.1871698>
- [20] Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural Text Generation from Structured Data with Application to the Biography Domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas, 1203–1213. <https://doi.org/10.18653/v1/D16-1128>
- [21] Adam Lerer, Ledell Wu, Jiajun Shen, Timothee Lacroix, Luca Wehrstedt, Abhijit Bose, and Alex Peysakhovich. 2019. PyTorch-BigGraph: A Large-scale Graph Embedding System. In *Proceedings of the 2nd SysML Conference*. Palo Alto, CA, USA.
- [22] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7871–7880. <https://doi.org/10.18653/v1/2020.acl-main.703>
- [23] Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by Summarizing Long Sequences. arXiv:1801.10198 [cs.CL]
- [24] Shan Liu and Mizuho Iwaihara. 2016. Extracting representative phrases from Wikipedia article sections. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*. 1–6. <https://doi.org/10.1109/ICIS.2016.7550850>
- [25] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics* 8 (2020), 726–742. https://doi.org/10.1162/tacla_a_00343
- [26] Jared K Lunceford and Marie Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 23 (2004).
- [27] Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and QUAN Z. Sheng. 2022. Multi-Document Summarization via Deep Learning Techniques: A Survey. *ACM Comput. Surv.* (mar 2022). <https://doi.org/10.1145/3529754> Just Accepted.
- [28] Kathleen McKeown and Dragomir R. Radev. 1995. Generating summaries of multiple news articles. In *SIGIR '95*.
- [29] Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. arXiv:1602.06023 [cs.CL]
- [30] 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:1808.08745 [cs.CL]
- [31] Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The Language Demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics* 2 (2014), 79–92. https://doi.org/10.1162/tacla_a_00167
- [32] María Pérez-Ortiz and Rafal K. Mantiuk. 2017. A practical guide and software for analysing pairwise comparison experiments. *ArXiv abs/1712.03686* (2017).
- [33] Maxime Peyrard, Wei Zhao, Steffen Eger, and Robert West. 2021. Better than Average: Paired Evaluation of NLP systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online, 2301–2315. <https://doi.org/10.18653/v1/2021.acl-long.179>
- [34] Tiziano Piccardi, Michele Catasta, Leila Zia, and Robert West. 2018. Structuring Wikipedia Articles with Section Recommendations. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) (SIGIR '18). Association for Computing Machinery, New York, NY, USA, 665–674. <https://doi.org/10.1145/3209978.3209984>
- [35] Christina Sauper and Regina Barzilay. 2009. Automatically Generating Wikipedia Articles: A Structure-Aware Approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapore, 208–216. <https://aclanthology.org/P09-1024>
- [36] Satoshi Sekine and Chikashi Nobata. 2003. A survey for Multi-Document Summarization. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*. 65–72. <https://aclanthology.org/W03-0509>
- [37] Tomás Sáez and Aidan Hogan. 2018. Automatically Generating Wikipedia Info-boxes from Wikidata. *WWW '18: Companion Proceedings of the The Web Conference 2018*, 1823–1830. <https://doi.org/10.1145/3184558.3191647>
- [38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.), 5998–6008. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fb0d053c1c4a845aa-Abstract.html>
- [39] Pavlos Vougiouklis, Hady Elsahar, Lucie-Aimée Kaffee, Christophe Gravier, Frédérique Laforest, Jonathon Hare, and Elena Simperl. 2018. Neural Wikipedian: Generating Textual Summaries from Knowledge Base Triples. *Journal of Web Semantics* 52–53 (2018), 1–15. <https://doi.org/10.1016/j.websem.2018.07.002>

- [40] Pavlos Vougiouklis, Eddy Maddalena, Jonathon S. Hare, and Elena Paslaru Bontas Simperl. 2020. Point at the Triple: Generation of Text Summaries from Knowledge Base Triples. *J. Artif. Intell. Res.* 69 (2020), 1–31.
- [41] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (sep 2014), 78–85. <https://doi.org/10.1145/2629489>
- [42] Denny Vrandečić. 2021. Building a multilingual Wikipedia. *Commun. ACM* 64, 4 (March 2021), 38–41. <https://doi.org/10.1145/3425778>
- [43] Roberto Yus, Varish Mulwad, Tim Finin, and Eduardo Mena. 2014. Infoboxer: Using Statistical and Semantic Knowledge to Help Create Wikipedia Infoboxes. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272* (Riva del Garda, Italy) (*ISWC-PD'14*). CEUR-WS.org, Aachen, DEU, 405–408.
- [44] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGA-SUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. [arXiv:1912.08777 \[cs.CL\]](https://arxiv.org/abs/1912.08777)
- [45] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 563–578. <https://doi.org/10.18653/v1/D19-1053>
- [46] Chenguang Zhu, William Hinthon, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing Factual Consistency of Abstractive Summarization. [arXiv:2003.08612 \[cs.CL\]](https://arxiv.org/abs/2003.08612)

A WIKIDATA DESCRIPTIONS

Wikipedia short description provide a concise indication of the field covered by the article. There exists a similar concept in Wikidata called *Wikidata descriptions*,⁷ whose purpose is to disambiguate Wikidata items with similar labels. One might be tempted to use Wikidata descriptions as Wikipedia descriptions and *vice versa*, but as Table 5 shows, this cannot solve the problem, as a significant fraction of Wikidata items have no description either—and when they do have a description, the corresponding Wikipedia article already does so, too, in the vast majority of cases. This high overlap is indicated by the Jaccard coefficient in the rightmost column of Table 5. Additionally, when both Wikipedia and Wikidata descriptions are present, they are almost always the same, with the exception of English, where only 58.37% are identical.

B ARTICLE FUSION

In order to fuse the article representations A_l obtained by encoding the article text with mBART encoder for each language $l \in \{1, \dots, n\}$ into a single representation A , we follow the process described in Sec. 3. More formally, the i -th token embedding A_i of A is calculated as

$$\frac{1}{n} \sum_{l=1}^n \text{FF} \left(\text{LayerNorm} \left(Q_i + \text{softmax} \left(\frac{Q_i K_l^T}{\sqrt{d}} \right) V_l \right) \right),$$

where $Q_i = (A_q)_i W_Q$, $K_l = A_l W_K$, $V_l = A_l W_V$. Vector $(A_q)_i$ represents the i -th token embedding of A_q . Matrices W_Q , W_K , W_V are trainable query, key, and value weights, and d is the dimension of the key. FF corresponds to the entire feed-forward network.

C ERROR TAXONOMY

We adopted an iterative coding strategy by manually inspecting the data to build an understanding of the types of error that occur in the descriptions. In each round, two authors independently labeled the same 100 randomly chosen samples, 50 where Descartes’s description was preferred, and 50 where human-written description

Table 5: Statistics of Wikidata items connected to 25 Wikipedia language editions.

Language	Articles	Missing desc.	Missing desc. (%)	Exact copies (%)	Jaccard	
en	English	5204K	556K	10.68	58.37	0.8904
de	German	2041K	321K	15.71	96.75	0.9585
nl	Dutch	1886K	181K	9.60	98.60	0.9913
es	Spanish	1463K	631K	43.17	95.16	0.9263
it	Italian	1287K	425K	32.99	96.22	0.9510
ru	Russian	1406K	877K	62.39	96.53	0.8413
fr	French	979K	237K	24.20	96.11	0.9164
zh	Chinese	1025K	836K	81.52	80.15	0.7830
ar	Arabic	986K	289K	29.31	98.96	0.9604
vi	Vietnamese	122K	811K	66.35	92.46	0.1226
ja	Japanese	1103K	768K	69.62	94.83	0.7297
fi	Finnish	451K	270K	59.95	97.48	0.8303
ko	Korean	422K	361K	85.58	94.76	0.7522
tr	Turkish	321K	232K	72.50	91.76	0.7584
ro	Romanian	282K	158K	56.29	93.38	0.9694
cs	Czech	178K	72K	40.19	96.62	0.8693
et	Estonian	195K	156K	79.74	99.51	0.8962
lt	Lithuanian	185K	175K	94.88	98.89	0.9512
kk	Kazakh	220K	219K	99.45	95.54	0.5842
lv	Latvian	92K	70K	76.30	96.35	0.9327
hi	Hindi	130K	75K	57.32	96.66	0.9043
ne	Nepali	29K	25K	84.57	94.74	0.9263
my	Burmese	44K	37K	84.46	95.02	0.8044
si	Sinhala	17K	15K	91.83	95.40	0.7110
gu	Gujarati	29K	7K	24.37	99.60	0.9827

was preferred, with the possibility to expand the set of labels (error categories) in each round. At the end of each round, the labelers discussed disagreements and the appropriateness of the selected categories, and modified them if needed. As a stopping criterion, an average pairwise Fleiss $\kappa > 0.6$ was used, and the coding process terminated after two rounds, with $\kappa = 0.55$ and 0.77 , respectively.

D ADDITIONAL RESULTS

In this section we present some of the examples of generated descriptions by our models. In Table 6, 7, and 8 we present some of the examples where semantic type and existing descriptions help.

⁷<https://www.wikidata.org/wiki/Help:Description>

Table 6: Examples where Descartes [no types] performs better than Descartes [no desc/types]

Article	Target	Descartes [no desc/types]	Descartes [no types].
Tony Leung Tony Leung Ka-Fai, nato nel 1958, chiamato anche "Big Tony" – attore cinese di Hong Kong	pagina di disambiguazione di un progetto Wikimedia	attore cinese	pagina di disambiguazione di un progetto Wikimedia
Tony Leung Chiu-Wai, nato nel 1962, chiamato anche "Little Tony" – attore cinese di Hong Kong			
Tony Leung Siu-Hung – attore, coreografo, regista e stuntman cinese di Hong Kong			
Tony Leung Hung-Wah – regista, produttore e sceneggiatore cinese di Hong Kong			
The United Nations Girls' Education Initiative (UNGEI) is an initiative launched by the United Nations in 2000 at the World Education Forum in Dakar at the primary school Ndiarème B. It aims to reduce the gender gap in schooling for girls and to give girls equal access to all levels of education.	organization	initiative launched by the United Nations in 2000	organization
De Golf Cup of Nations 1974 was de 3e editie van dit voetbaltoernooi dat werd gehouden in Koeweit van 15 maart 1974 tot en met 29 maart 1974. Koeweit won het toernooi door in de finale Saoedi-Arabië te verslaan.	sportseizoen van een voetbalcompetitie	golftournooi	sportseizoen van een voetbalcompetitie
Peter DeGraaf is a Republican member of the Kansas House of Representatives, representing the 82nd district. He has served since May 2008.	American politician	Kansas House of Representatives	American politician
The Tree of Life (Shajarat-al-Hayat) in Bahrain is a 9.75 meters high <i>Prosopis cineraria</i> tree that is over 400 years old. It is on a hill in a barren area of the Arabian Desert, 2 kilometers from Jebel Dukhan, the highest point in Bahrain, and 40 kilometers from Manama.	tree in Bahrain	species of plant	tree in Bahrain

Table 7: Examples where Descartes [no desc] performs better than Descartes [no desc/types]

Article	Target	Descartes [no desc/types]	Descartes [no desc]
Corticotropin-releasing hormone (CRH) is a peptide hormone involved in the stress response. It is a releasing hormone that belongs to corticotropin-releasing factor family. In humans, it is encoded by the CRH gene. Its main function is the stimulation of the pituitary synthesis of ACTH, as part of the HPA Axis.	mammalian protein found in Homo sapiens	chemical compound	mammalian protein found in Homo sapiens
Postcrossing is een internationaal briefkaart-uitwisselingsproject. Het werd op 13 juli 2005 opgezet door de uit Portugal afkomstige Paulo Magalhães. Postcrossing is nadien uitgegroeid tot een uitwisselingsproject met ruim 800.000 gebruikers uit 206 verschillende landen. In februari 2017 werd de 40.000.000ste postkaart die via postcrossing werd verstuurd ontvangen.	website	organisatie uit Portugal	website
Ligne Namboku (南北線, Namboku-sen), littéralement ligne Sud-Nord, est le nom donné à plusieurs lignes ferroviaires au Japon: la ligne Namboku du métro de Tokyo; la ligne Namboku du métro de Sendai; la ligne Namboku du métro de Sapporo; la ligne Namboku de la compagnie Kita-Osaka Kyuko Railway; la ligne Namboku de la compagnie Kobe Rapid Transit Railway exploitée par la Kobe Electric Railway sous le nom de ligne Kobe Kosoku.	page d'homonymie de Wikimedia	ligne de chemin de fer japonaise	page d'homonymie de Wikimedia
The Atari Jaguar is a 64-bit home video game console developed by Atari Corporation and designed by Flare Technology, released in North America first on November 23, 1993. It was the sixth programmable console developed under the Atari brand. The following list contains all of the games released on cartridge for the Jaguar.	Wikipedia list article	1993 video game console	Wikipedia list article
The 2017 BWF World Junior Championships was the nineteenth tournament of the BWF World Junior Championships. It was held in Yogyakarta, Indonesia at the Among Rogo Sports Hall between 9 and 22 October 2017.	badminton championships	2017 edition of the World Junior Championships	badminton championships

Table 8: Examples where Descartes performs better than other variations of Descartes

Article	Target	Descartes [no desc/types]	Descartes [no]	Descartes [no types]	Descartes [no desc]	Descartes
Het Europees kampioenschap volleybal mannen 2003 vond van 5 tot en met 14 september plaats in Karlsruhe en Leipzig (Duitsland).	sportseizoen van een volleybalcompetitie	Volleyball-Europa	mannenekelspel	mannenekelspel	sportseizoen van een volleybalcompetitie	
Mohamed Shies Madhar is een Surinaams voormalig judoka.	judoka uit Suriname	Nederlands judoka	Nederlands judoka	judoka	judoka uit Suriname	
Guanosine monophosphate synthetase, also known as GMPS is an enzyme that converts xanthosine monophosphate to guanosine monophosphate.	mammalian protein found in Homo sapiens	class of enzymes	class of enzymes	class of enzymes	mammalian protein found in Homo sapiens	
La gmina de Lubowidz est un district administratif situé en milieu rural du powiat de Żuromin dans la voïvodie de Mazovie, dans le centre-est de la Pologne.	commune polonaise	gmina rurale polonaise	gmina rurale polonaise	gmina rurale polonaise	commune polonaise	
Membrane-bound transcription factor site-1 protease, or site-1 protease (S1P) for short, also known as subtilisin/kexin-isozyme 1 (SKI-1), is an enzyme that in humans is encoded by the MBTPS1 gene. S1P cleaves the endoplasmic reticulum loop of sterol regulatory element-binding protein (SREBP) transcription factors.	mammalian protein found in Homo sapiens	protein-coding gene in the species Homo sapiens	protein-coding gene in the species Homo sapiens	protein in Homo sapiens	mammalian protein found in Homo sapiens	