

Using natural language generation to bootstrap missing Wikipedia articles: A human-centric perspective

Lucie-Aimée Kaffee^{a,*}, Pavlos Vougiouklis^{b,**} and Elena Simperl^c

^a *School of Electronics and Computer Science, University of Southampton, UK*

E-mail: kaffee@soton.ac.uk

^b *Huawei Technologies, UK*

E-mail: pavlos.vougiouklis@huawei.com

^c *King's College London, UK*

E-mail: elena.simperl@kcl.ac.uk

Editor: Philipp Cimiano, Universität Bielefeld, Germany

Solicited reviews: John Bateman, Bremen University, Germany; Leo Wanner, Pompeu Fabra University, Spain; Denny Vrandečić, Wikimedia Foundation, USA

Abstract. Nowadays natural language generation (NLG) is used in everything from news reporting and chatbots to social media management. Recent advances in machine learning have made it possible to train NLG systems that seek to achieve human-level performance in text writing and summarisation. In this paper, we propose such a system in the context of Wikipedia and evaluate it with Wikipedia readers and editors. Our solution builds upon the ArticlePlaceholder, a tool used in 14 under-resourced Wikipedia language versions, which displays structured data from the Wikidata knowledge base on empty Wikipedia pages. We train a neural network to generate an introductory sentence from the Wikidata triples shown by the ArticlePlaceholder, and explore how Wikipedia users engage with it. The evaluation, which includes an automatic, a judgement-based, and a task-based component, shows that the summary sentences score well in terms of perceived fluency and appropriateness for Wikipedia, and can help editors bootstrap new articles. It also hints at several potential implications of using NLG solutions in Wikipedia at large, including content quality, trust in technology, and algorithmic transparency.

Keywords: Wikipedia, Wikidata, ArticlePlaceholder, multilingual, natural language generation, neural networks

1. Introduction

Wikipedia is available in 301 languages, but its content is unevenly distributed [31]. Language versions with less coverage than e.g. English Wikipedia face multiple challenges: fewer editors means less quality control, making that particular Wikipedia less attractive for readers in that language, which in turn makes it more difficult to recruit new editors from among the readers.

*Corresponding author. E-mail: kaffee@soton.ac.uk.

**Work done while working at University of Southampton.

Saint-Marcellin (Q510126)		saint-marcellin (Q510126)		Saint-marcellin	
subclass of (P279)	cow's-milk cheese (Q3088299) white mold-rind cheese (Q1256296) farmstead cheese (Q3088318) dairy cheese (Q16637313)	sous-classe de (P279)	fromage au lait de vache (Q3088299) fromage à pâte molle à croûte fleurie (Q1256296) fromage fermier (Q3088318) fromage laitier (Q16637313)		
image (P18)	Wikicheese - Saint-marcellin - 20150417 - 010.jpg	image (P18)	Wikicheese - Saint-marcellin - 20150417 - 010.jpg		
country of origin (P495)	France (Q142)	pays d'origine (P495)	France (Q142)	Pays d'origine	France ✓
material used (P186)	cow milk (Q10988133) applies to part, aspect, or form: milk (P518: Q8495)	matériau (P186)	lait de vache (Q10988133) s'applique à: lait (P518: Q8495)	Région	Isère ✓
location of creation (P1071)	Isère (Q12559)	lieu de fabrication (P1071)	Isère (Q12559)	Lait	Vache ✓
product certification (P1389)	Geographical indications and traditional specialities of the European Union (Q363268)	certificat de produit (P1389)	Indications géographiques protégées de l'Union Européenne (Q363268)	Pâte	Molle à croûte fleurie ✓
reference	http://www.fromage-saint-marcellin.fr/des-savoir-faire-sacres/qualite-74.html	référence	http://www.fromage-saint-marcellin.fr/des-savoir-faire-sacres/qualite-74.html	Appellation	Indications géographiques protégées de l'Union Européenne (en) (2013) ✓
				Nommé en référence à	Saint-Marcellin ✓

Fig. 1. Representation of Wikidata statements and their inclusion in a Wikipedia infobox. Wikidata statements in French (middle, English translation to their left) are used to fill out the fields of the infobox in articles using the *fromage* infobox on the French Wikipedia.

Wikidata, the structured-data backbone of Wikipedia [86], offers some help. It contains information about more than 55 million entities, for example, people, places or events, edited by an active international community of volunteers [40]. More importantly, it is multilingual by design and each aspect of the data can be translated and rendered to the user in their preferred language [39]. This makes it the tool of choice for a variety of content integration affordances in Wikipedia, including links to articles in other languages and infoboxes. An example can be seen in Fig. 1: in the French Wikipedia, the infobox shown in the article about cheese (right) automatically draws in data from Wikidata (left) and displays it in French.

In previous work of ours, we proposed the ArticlePlaceholder, a tool that takes advantage of Wikidata's multilingual capabilities to increase the coverage of under-resourced Wikipedias [41]. When someone looks for a topic that is not yet covered by Wikipedia in their language, the ArticlePlaceholder tries to match the topic with an entity in Wikidata. If successful, it then redirects the search to an automatically generated *placeholder page* that displays the relevant information, for example the name of the entity and its main properties, in their language. The ArticlePlaceholder is currently used in 14 Wikipedias (see Section 3.1).

In this paper, we propose an iteration of the ArticlePlaceholder to improve the representation of the data on the placeholder page. The original version of the tool pulled the raw data from Wikidata (available as triples with labels in different languages) and displayed it in tabular form (see Fig. 3 in Section 3). In the current version, we use Natural Language Generation (NLG) techniques to automatically produce a single summary sentence from the triples instead. Presenting structured data as text rather than tables helps people uninitiated with the involved technologies to make sense of it [84]. This is particularly useful in contexts where one cannot make any assumptions about the levels of data literacy of the audience, as is the case for a large share of the Wikipedia readers.

Our NLG solution builds upon the general *encoder-decoder framework* for neural networks, which is credited with promising results in similar text-centric tasks, such as machine translation [12,82] and question generation [16,19,78]. We extend this framework to meet the needs of different Wikipedia language communities in terms of text fluency, appropriateness to Wikipedia, and reuse during article editing. Given an entity that was matched by the ArticlePlaceholder, our system uses its triples to generate a Wikipedia-style summary sentence. Many existing NLG techniques produce sentences with limited usability in user-facing systems; one of the most common problems is their ability to handle rare words [15,58], which are words that the model does not meet frequently enough during training, such as localisations of names in different languages. We introduce a mechanism called *property placeholder* [36] to tackle this problem, learning multiple verbalisations of the same entity in the text [84].

In building the system we aimed to pursue the following research questions:

RQ1 Can we train a neural network to generate text from triples in a multilingual setting? To answer this question we first evaluated the system using a series of predefined metrics and baselines. In addition, we undertook a quantitative study with participants from two different Wikipedia language communities (Arabic

and Esperanto), who were asked to assess, from a reader's perspective, whether the sentence is fluent and appropriate for Wikipedia.

RQ2 How do editors perceive the generated text on the ArticlePlaceholder page? To add depth to the quantitative findings of the first study, we undertook a second, mixed-methods study within six Wikipedia language communities (Arabic, Swedish, Hebrew, Persian, Indonesian, and Ukrainian). We carried out semi-structured interviews, in which we asked editors to comment on their experience with reading the summary sentences generated through our approach and we identified common themes in their answers. Among others, we were interested to understand how editors perceive text that is the result of the artificial intelligence (AI) algorithm rather than being manually crafted, and how they deal with so-called `<rare>` tokens in the sentences. Those tokens represent realisations of infrequent entities in the text, that data-driven approaches generally struggle to verbalise [58].

RQ3 How do editors use the generated sentence in their work? As part of the second study, we also asked participants to edit the placeholder page, starting from the automatically generated text or removing it completely. We assessed text reuse both quantitatively, using a string-matching metric, and qualitatively through the interviews. Just like in RQ2, we were also interested to understand whether summaries with `<rare>` tokens, which point to limitations in the algorithm, would be used when editing and how the editors would work around the tokens.

The evaluation helps us build a better understanding of the tools and experience we need to help nurture under-served Wikipedias. Our quantitative analysis of the reading experience showed that participants rank the summary sentences close to the expected quality standards in Wikipedia, and are likely to consider them as part of Wikipedia. This was confirmed by the interviews with editors, which suggested that people believe the summaries to come from a Wikimedia-internal source. According to the editors, the new format of the ArticlePlaceholder enhances the reading experience: people tend to look for specific bits of information when accessing a Wikipedia page and the compact nature of the generated text supports that. In addition, the text seems to be a useful starting point for further editing and editors reuse a large portion of it even when it includes `<rare>` tokens.

We believe the two studies could also help advance the state of the art in two other areas: together, they propose a user-centred methodology to evaluate NLG, which complements automatic approaches based on standard metrics and baselines, which are the norm in most papers; at the same time, they also shed light on the emerging area of human-AI interaction in the context of NLG. While the editors worked their way around the `<rare>` tokens both during reading and writing, they did not check the text for correctness, nor queried where the text came from and what the tokens meant. This suggests that we need more research into how to communicate the provenance of content in Wikipedia, especially in the context of automatic content generation and deep fakes [34], as well as algorithmic transparency.

Structure of the paper The remainder of the paper is organised as follows. We start with some background for our work and related papers in Section 2. Next, we introduce our approach to bootstrapping empty Wikipedia articles, which includes the ArticlePlaceholder tool and its NLG extension (Section 3). In Section 4 we provide details on the evaluation methodology, whose findings we present in Section 5. We then discuss the main themes emerging from the findings and their implications and the limitations of our work in Section 6 and 7, before concluding with a summary of contributions and planned future work in Section 8.

Previous submissions A preliminary version of this work was published in [36,37]. In the current paper, we have carried out a comprehensive evaluation of the approach, including a new qualitative study and a task-based evaluation with editors from six language communities. By comparison, the previous publications covered only a metric-based corpus evaluation which was complemented by a small quantitative study of text fluency and appropriateness in the second one. The neural network architecture has been presented in detail in [36].

2. Background and related work

We divide this section into three areas. First we provide some background on Wikipedia and Wikidata, with a focus on multilingual aspects. Then we summarise the literature on text generation and conclude with methodologies to evaluate NLG systems.

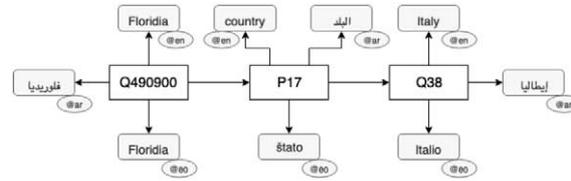


Fig. 2. Example of labeling in Wikidata. Each entity can be labeled in multiple languages using the labeling property `rdfs:label` (not displayed in the figure).

2.1. Multilingual Wikipedia and Wikidata

Wikipedia is a community-built encyclopedia and one of the most visited websites in the world. There are currently 301 language versions of Wikipedia, though coverage is unevenly distributed. Previous studies have discussed several biases, including gender of the editors [14], and topics, for instance a general lack of information on the Global South [24].

Language coverage tells a similar story. [67] noted that only 67% of the world’s population has access to encyclopedic knowledge in their first or second language. Another significant problem is caused by the extreme differences in content coverage between language versions. As early as 2005, Voss [83] found huge gaps in the development and growth of Wikipedias, which make it more difficult for smaller communities to catch up with the larger ones. Alignment cannot be simply achieved through translation – putting aside the fact that each Wikipedia needs to reflect the interests and points of view of their local community rather than iterate over content transferred from elsewhere, studies have shown that the existing content does not overlap, discarding the so-called *English-as-Superset* conjecture for as many as 25 language versions [31,32]. To help tackle these imbalances, Bao et al. [6] built a system to give users easier access to the language diversity of Wikipedia.

Our work is motivated and complements previous studies and frameworks that argue that the language of global projects such as Wikipedia [32] should express cultural reality [46]. Instead of using content from one Wikipedia version to bootstrap another, we take structured data labelled in the relevant language and create a more accessible representation of it as text, keeping those cultural expressions unimpaired in terms of language composition compared to machine translation.

To do so, we leverage Wikidata [86]. Wikidata was originally created to support Wikipedia’s language connections, for instance links to articles in other languages or infoboxes, see example from Section 1); however, it soon evolved into becoming a critical source of data for many other applications.

Wikidata contains statements on general knowledge, e.g. about people, places, events and other entities of interest. The knowledge base is created and maintained collaboratively by a community of editors, assisted by automated tools called bots [38,80]. Bots take on repetitive tasks such as ingesting data from a different source, or simple quality checks.

The basic building blocks of Wikidata are *items* and *properties*. Both have identifiers, which are language-independent, and labels in one or more languages. They form triples, which link items to their attributes, or to other items. Figure 2 shows a triple linking a place to its country, where both items and the property between them have labels in three languages.

We analysed the distribution of label languages in Wikidata in [39] and noted that while there is a bias to English, the language distribution is more balanced than on the web at large. This was the starting point for our work on the ArticlePlaceholder (see Section 3), which leverages the multilingual support of Wikidata to bootstrap empty articles in less-resourced Wikipedias.

2.2. Text generation

Our works builds on top of prior research that looked at generating texts, without intermediate Machine Translation stages, preserving the cultural characteristics of the target languages [7,29,42,47]. In our task, we focus on generating sentences from triples expressed as Resource Description Framework (RDF) or similar. Many of the related approaches rely on templates, which are either based on linguistic features e.g., grammatical rules [87] or

are hand-crafted [20]. An example is *Reasonator*, a tool for lexicalising Wikidata triples with templates translated by users.¹ Such approaches face many challenges – they cannot be easily transferred to different languages or scale to broader domains, as templates need to be adjusted to any new language or domain they are ported to. This makes them unsuitable for Wikipedias which rely on small numbers of contributors.

To tackle this limitation, Duma et al. [17] and Ell et al. [18] introduced a distant-supervised method to verbalise triples, which learns templates from existing Wikipedia articles. While this makes the approach more suitable for language-independent tasks, templates assume that entities will always have the relevant triples to fill in the slots. This assumption is not always true. In our work, we propose a template-learning baseline and show that adapting to the varying triples available can achieve better performance.

Sauper and Barzilay [76] and Pochampally et al. [69] generate Wikipedia summaries by harvesting sentences from the Internet. Wikipedia articles are used to automatically derive templates for the topic structure of the summaries and the templates are afterward filled using web content. Both systems work best on specific domains and for languages like English, for which suitable web content is readily available [53].

There is a large body of work that uses the encoder-decoder framework from machine translation [12,82] for NLG [11,21,22,51,55,59,79,84,90,91]. Adaptations of this framework have shown great potential at tackling various aspects of triples-to-text tasks ranging from microplanning by Gardent et al. [21] to generation of paraphrases by Sleimi and Gardent [79]. Mei et al. [59] sought to generate textual descriptions from datasets related to weather forecasts and RoboCup football matches. Wiseman et al. [90] used pointer-generator networks [77] to generate descriptions of basketball games, while Gehrmann et al. [22] did the same for restaurant descriptions.

A different line of research aims to explore knowledge bases as a resource for NLG [10,11,17,51,55,84,91]. In all these examples, linguistic information from the knowledge base is used to build a parallel corpus containing triples and equivalent text sentences from Wikipedia, which is then used to train the NLG algorithm. Directly relevant to the model we propose are the proposals by Lebret et al. [51], Chisholm et al. [11], Liu et al. [55], Yeh et al. [91] and Vougiouklis et al. [84,85], which extend the general encoder-decoder neural network framework from [12,82] to generate short summaries in English. The original of English biographies generation was introduced by Lebret et al. [51] who used feed-forward language model with slot-value templates to generate the first sentence of a Wikipedia summary from its corresponding infobox. Incremental upgrades of the original architecture on the same task include the introduction of an auto-encoding pipeline based on an attentive encoder-decoder architecture using GRUs [11], a novel double-attention mechanism over the input infoboxes' fields and their values [55], and adaptations of pointer-generator mechanisms [85,91] over the input triples.

All these approaches use structured data from Freebase, Wikidata and DBpedia as input and generate summaries consisting either by one or two sentences that match the style of the English Wikipedia in a single domain [11,51,55,84,91] or more recently in more open-domain scenarios [85]. While this is a rather narrow task compared to other generative tasks such as translation, Chisholm et al. [11] discuss its challenges in detail and show that it is far from being solved. Compared to these previous works, the NLG algorithm presented in this paper is for open-domain tasks and multiple languages.

2.3. Evaluating text generation systems

Related literature suggests three ways of determining how well an NLG system achieves its goals. The first, which is commonly referred to as *metric-based corpus evaluation* [72], use text-similarity metrics such as BLEU [66], ROUGE [54] and METEOR [50]. These metrics essentially compare how similar the generated texts are to texts from the corpus. The other two involve people and are either *task-based* or *judgement/rating-based* [72]. Task-based evaluations aim to assess how an NLG solution assists participants in undertaking a particular task, for instance learning about a topic or writing a text. judgement-based evaluations rely on a set of criteria against which participants are asked to rate the quality of the automatically generated text [72].

Metric-based corpus evaluations are widely used as they offer an affordable, reproducible way to automatically assess the linguistic quality of the generated texts [4,11,36,45,51,72]. However, they do not always correlate with manually curated quality ratings [72].

¹<https://tools.wmflabs.org/reasonator/>

Task-based studies are considered most useful, as they allow system designers to explore the impact of the NLG solution to end-users [60,72]. However, they can be resource-intensive – previous studies [73,88] cite five figure sums, including data analysis and planning [71]. The system in [88] was evaluated for the accuracy of the generated literacy and numeracy assessments by a sample of 230 participants, which cost as much as 25 thousand UK pounds. Reiter et al. [73] described a clinical trial with over 2000 smokers costing 75 thousand pounds, assumed to be the most costly NLG evaluation at this point. All of the smokers completed a smoking questionnaire in the first stage of the experiment, in order to find what portion of those who received the automatically generated letters from STOP had managed to quit.

Given these challenges, most research systems tend to use judgement-based rather than task-based evaluations [4,11,17,18,45,64,72,81]. Besides their limited scope, most studies in this category do not recruit from the relevant user population, relying on more accessible options such as online crowdsourcing. Sauper and Barzilay [76] is a rare exception. In their paper, the authors describe the generation of Wikipedia articles using a content-selection algorithm that extracts information from online sources. They test the results by publishing 15 articles about diseases on Wikipedia and measuring how the articles change (including links, formatting and grammar). Their evaluation approach is not easy to replicate, as the Wikipedia community tends to disagree with conducting research on their platform.²

The methodology we follow draws from all these three areas: we start with an automatic, metric-based corpus evaluation to get a basic understanding of the quality of the text the system produces and compare it with relevant baselines. We then use quantitative analysis based on human judgements to assess how useful the summary sentences are for core tasks such as reading and editing. To add context to these figures, we run a mixed-methods study with an interview and a task-based component, where we learn more about the user experience with NLG and measure the extent to which NLG text is reused by editors, using a metric inspired from journalism [13] and plagiarism detection [70].

3. Bootstrapping empty Wikipedia articles

The overall aim of our system is to give editors access to information that is not yet covered in Wikipedia, but is available, in the relevant language, in Wikidata. The system is built on the ArticlePlaceholder that displays Wikidata triples dynamically on different language Wikipedias. In this paper, we extend the ArticlePlaceholder with an NLG component that generates an introductory sentence on each ArticlePlaceholder page in the target language from Wikidata triples.

3.1. ArticlePlaceholder

As discussed earlier, some Wikipedias suffer from a lack of content, which means fewer readers, and in turn, fewer potential editors. The idea of the ArticlePlaceholder is to use Wikidata, which contains information about 55 million entities (by comparison, the English Wikipedia covers around 5 million topics), often in different languages, to bootstrap articles in language versions lacking content. An initial version of this tool was presented in [41].

ArticlePlaceholders are pages on Wikipedia that are dynamically drawn from Wikidata triples. When the information in Wikidata changes, the ArticlePlaceholder pages are automatically updated. In the original release, the pages display the triples in a tabular way, purposely not reusing the design of a standard Wikipedia page to make the reader aware that the page was automatically generated and requires further attention. An example of the interface can be seen in Fig. 3.

The Article Placeholder is deployed on 14 Wikipedias with a median of 69,623.5 articles, between 253,539 (Esperanto) and 7,464 (Northern Sami).

²https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not#Wikipedia_is_not_a_laboratory

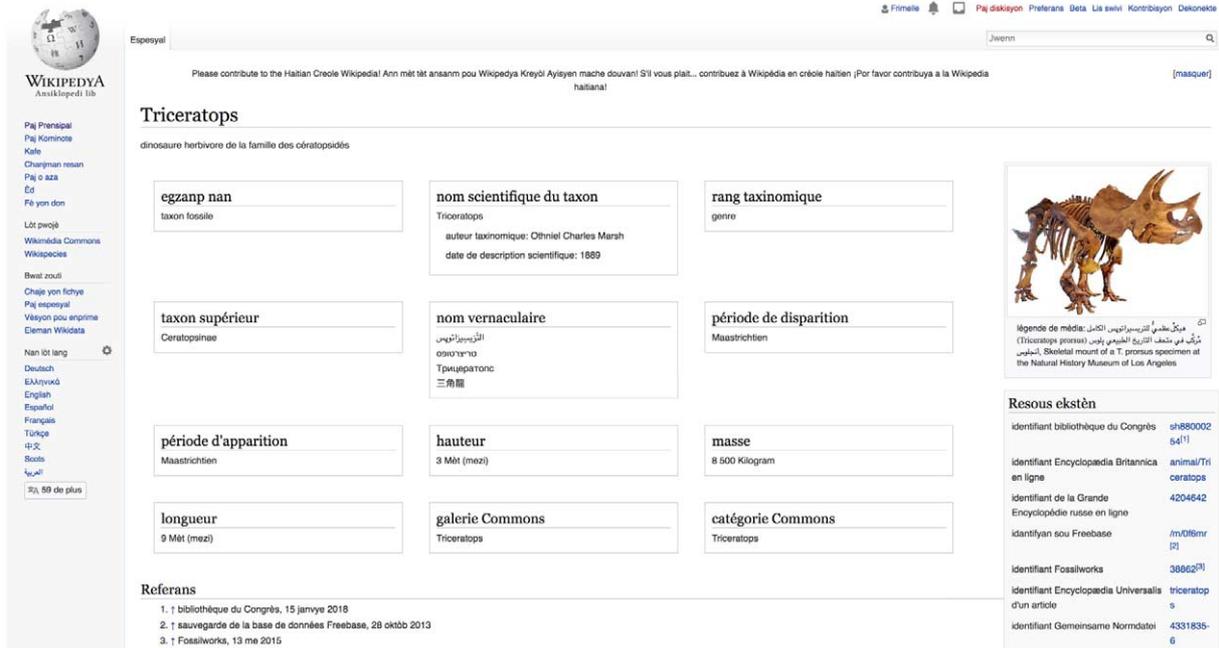


Fig. 3. Example page of the ArticlePlaceholder as deployed now on 14 Wikipedias. This example contains information from Wikidata on Triceratops in Haitian-Creole.

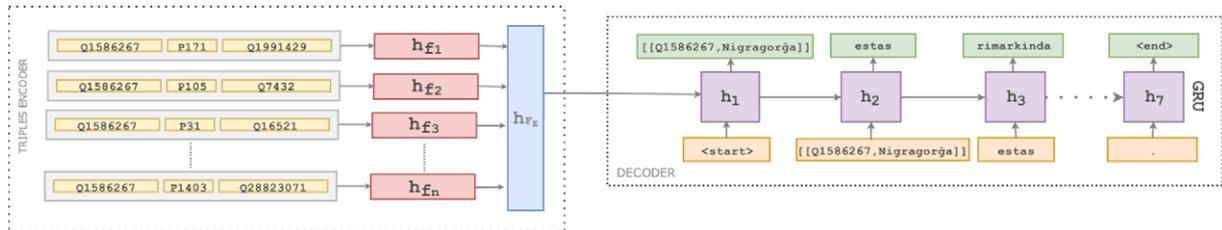


Fig. 4. Representation of the neural network architecture. The triple encoder computes a vector representation for each one of the three input triples from the ArticlePlaceholder, h_{f_1} , h_{f_2} and h_{f_3} . Subsequently, the decoder is initialised using the concatenation of the three vectors, $[h_{f_1}; h_{f_2}; h_{f_3}]$. The purple boxes represent the tokens of the generated text. Each snippet starts and ends with special tokens: start-of-summary $\langle \text{start} \rangle$ and end-of-summary $\langle \text{end} \rangle$. Example in Esperanto.

3.2. Text generation

We use a data-driven approach that allows us to extend the ArticlePlaceholder pages with a short description of the article’s topic.

3.2.1. Neural architecture

We reuse the encoder-decoder architecture introduced in previous work of ours in Vougiouklis et al. [84], which was focused on a closed-domain text generative task for English. The model consists of a feed-forward architecture, the *triple encoder*, which encodes an input set of triples into a vector of fixed dimensionality, and a Recurrent Neural Network (RNN) based on Gated Recurrent Units (GRUs) [12], which generates a sentence by conditioning the output on the encoded vector.

The model is displayed in Fig. 4. The ArticlePlaceholder provides a set of triples about the Wikidata item of *Nigragorĝa* (i.e., Q1586267 (Nigragorĝa) is either the subject or the object of the triples in the set). Figure 4 shows how the model generates a summary from those triples, f_1 , f_2 to f_n . A vector representation h_{f_1} , h_{f_2} to h_{f_n} for each of the input triples is computed by processing their subject, predicate and object.

Table 1

The ArticlePlaceholder provides our system with a set of triples about *Florida*, whose either subject or object is related to the item of Florida. Subsequently, our system summarizes the input set of triples as text. We train our model using the summary with the extended vocabulary (i.e. “Summary w/ Property placeholders”)

ArticlePlaceholder triples	f_1 :	Q490900 (Florida)	P17 (ŝtato)	Q38 (Italio)
	f_2 :	Q490900 (Florida)	P31 (estas)	Q747074 (komunumo de Italio)
	f_3 :	Q30025755 (Florida)	P1376 (ĉefurbo de)	Q490900 (Florida)
Textual summary		Florida estas komunumo de Italio.		
Summary w/ Property placeholders		[[Q490900, Florida]] estas komunumo de [[P17]].		

Such triple-level vector representations help compute a vector representation for the whole input set h_{FE} . h_{FE} , along with a special start-of-summary <start> token. This initialises the *decoder*, which predicts tokens (“[[Q1586267, Nigragorĝa]]”, “estas”, “rimarkinda” etc. for the generation in Esperanto).

Surface form tuples A summary such as “Nigragorĝa estas rimarkinda ...” consists of regular words (e.g. “estas” and “rimarkinda”) and mentions of entities in the text (“[[Q1586267, Nigragorĝa]]”). However, an entity can be expressed in a number of different ways in the summary. For example, “actor” and “actress” refer to the same concept in the knowledge graph. In order to be able to learn an arbitrary number of different lexicalisations of the same entity in the summary (e.g. “aktorino”, “aktoro”), we adapt the concept of *surface form tuples* [84]. “[Q1586267, Nigragorĝa]” is an example of a surface form tuple whose first element is the entity in the knowledge graph and its second is its realisation *in the context* of this summary. Any other surface form tuple for Q1586267 would have its first part identical to [[Q1586267, Nigragorĝa]] while its second part would include a different realisation. At a post processing step, when a token corresponding to a particular surface form tuple is generated by our model, we retain only the second part of the tuple in the text.

Property placeholders The model by Vougiouklis et al. [84], which was the starting point for the component presented here, leverages instance-type-related information from DBpedia in order to generate text that covers rare or unseen entities. We broadened its scope and adapted it to Wikidata without using external information from other knowledge bases. Some of the surface form tuple correspond to a rare entity, i.e., entities less frequently used and therefore potential out of vocabulary words in training. Given such a surface form tuple that is part of a triple matched to a sentence, we can leverage the relationship (or property) of this triple to use it as a special token to replace the rare surface form. The surface form tuple is replaced by the token of the property that matches the relationship. We refer to those placeholder tokens [15,78] as *property placeholders*. These tokens are appended to the target dictionary of the generated summaries. We used the distant-supervision assumption for relation extraction [61] for the property placeholders. After identifying the rare entities that participate in relations with the main entity of the article, they are replaced from the introductory sentence with their corresponding property placeholder tag (e.g. [[P17]] in Table 1). During testing, any property placeholder token that is generated by our system is replaced by the label of the entity of the relevant triple (i.e. triple with the same property as the generated token). In Table 1, [[P17]] in the processed summary is an example of a property placeholder. In case it is generated by our model, it is replaced with the label of the object of the triple with which they share the same property (i.e. Q490900 (Florida) P17 (ŝtato) Q38 (Italio)). We can show the impact of the property placeholders by measuring how many of the existing Wikipedia summaries we could recreate with the vocabulary used in training and test. In Esperanto, we could generate 3.7% of the sentences of the training and validation set, 8.7% in Arabic. Using property placeholders, we could generate 44.5% of the sentences of the training and validation set in Esperanto, 30.4% in Arabic.

3.2.2. Training dataset

In order to train and evaluate our system, we created a dataset for text generation from knowledge base triples in two languages. We used two language versions of Wikipedia (we provide further details about how we prepared the summaries for the rest of the languages in Section 4.3) which differ in terms of size (see Table 2) and language support in Wikidata [39]. The dataset aligns Wikidata triples about an item with the first sentence of the Wikipedia article about that entity.

Table 2

Page statistics and number of unique words (vocabulary size) of Esperanto, Arabic and English Wikipedias in comparison with Wikidata. Retrieved 27 September 2017. Active users are registered users that have performed an action in the last 30 days

Page stats	Esperanto	Arabic	English	Wikidata
Articles	241,901	541,166	5,483,928	37,703,807
Average number of edits/page	11.48	8.94	21.11	14.66
Active users	2,849	7,818	129,237	17,583
Vocabulary size	1.5M	2.2M	2M	–

For each Wikipedia article, we extracted and tokenized the first sentence using a multilingual Regex tokenizer from the NLTK toolkit [8]. Afterwards, we retrieved the corresponding Wikidata item to the article and queried all triples where the item appeared as a subject or an object in the Wikidata truthy dump.³

To map the triples to the extracted Wikipedia sentence, we relied on keyword matching against labels from Wikidata from the corresponding language, due to the lack of reliable entity linking tools for lesser resourced languages. For example, in the Esperanto sentence “*Floridia estas komunumo de Italio.*” (English: “*Floridia is a municipality of Italy.*”) for the Wikipedia article of *Floridia*, we extract all triples which have the Wikidata entity of *Floridia* as either subject or object. Then, we match “*Floridia*” in the sentence to the Wikidata entity *Floridia* (*Q490900*)⁴ based on the Wikidata Esperanto label.

We use property placeholders (as described in the previous section) to avoid the lack of vocabulary typical for under-resourced languages. An example of a summary which is used for the training of the neural architecture is: “*Floridia estas komunumo de [[P17]].*”, where [[P17]] is the property placeholder for *Italio* (see Table 1). In case a rare entity in the text is not matched to any of the input triples, its realisation is replaced by the special <rare> token.

4. Evaluation methodology

We followed a mixed-methods approach to investigate the three questions discussed in the introduction (Table 3). To answer *RQ1*, we used an automatic metric-based corpus evaluation, as well as a judgements-based quantitative evaluation with readers in two language Wikipedias, who are native (or fluent, in the case of Esperanto) speakers of the language.⁵ We showed them text generated through our approach, as well as genuine Wikipedia sentences and news snippets of similar length, and asked them to rate their fluency and appropriateness for Wikipedia on a scale. To answer *RQ2* and *RQ3*, we carried out an interview study with editors of six Wikipedias, with qualitative and quantitative components. We instructed editors to complete a reading and an editing task and to describe their experiences along a series of questions. We used thematic analysis to identify common themes in the answers. For the editing task, we also used a quantitative element in the form of a text reuse metric, which is described below.

4.1. *RQ1 – Metric-based corpus evaluation*

We evaluated the generated summaries against two baselines on their original counterparts from Wikipedia. We used a set of evaluation metrics for text generation BLEU 2, BLEU 3, BLEU 4, METEOR and ROUGE_L. BLEU calculates n-gram precision multiplied by a brevity penalty, which penalizes short sentences to account for word recall. METEOR is based on the combination of uni-gram precision and recall, with recall weighted over precision. It extends BLEU by including stemming, synonyms and paraphrasing. ROUGE_L is a recall-based metric which calculates the length of the most common subsequence between the generated summary and the reference.

³<https://dumps.wikimedia.org/wikidatawiki/entities/>

⁴<https://www.wikidata.org/wiki/Q490900>

⁵The raw data of the quantitative evaluation experiments can be found here: <https://github.com/pvougjou/Mind-the-Language-Gap/tree/master/crowdevaluation>.

Table 3
Evaluation methodology

	Data	Method	Participants
RQ1	Metrics and survey answers	Metric-based evaluation and judgement-based evaluation quantitative	Readers of two Wikipedias
RQ2	Interview answers	Task-based evaluation, qualitative (thematic analysis)	Editors of six Wikipedias
RQ3	Interview answers and text reuse metrics	Task-based evaluation, qualitative (thematic analysis) and quantitative	Editors of six Wikipedias

Table 4
Statistics of the two corpora. Average parameters are shown with standard deviations in brackets

Parameter	Arabic	Esperanto
Total # of Articles	255741	126714
Total # of Entities	355k	352k
Total # of Predicates	1021	965
Avg. # of Triples (incl. Encoded Dates) per Article	8.10 (11.23)	11.23 (13.82)
Max. # of Alloc. Triples (incl. Encoded Dates) per Article	885	883
Avg. # of Tokens per Summary	27.98 (28.57)	26.36 (22.71)
Total # of Words In the Summaries	433k	324k
Total # of Annotated Entities In the Summaries (excl. the Main Entity <item>)	22k	18k

4.1.1. Data

Both the Arabic and Esperanto corpus are split into training, validation and test, with respective portions of 85%, 10%, and 5%. We used a fixed target vocabulary that consisted of the 25000 and 15000 of the most frequent tokens (i.e. either words or surface form tuples of entities) of the summaries in Arabic and Esperanto respectively. Table 4 shows an overview of both languages and their datasets.

We replaced any rare entities in the text that participate in relations in the aligned triple set with the corresponding property placeholder of the upheld relations. We include all property placeholders that occur at least 20 times in each training dataset. Subsequently, the dictionaries of the Esperanto and Arabic summaries are expanded by 80 and 113 property placeholders respectively. In case the rare entity is not matched to any subject or object of the set of corresponding triples, it is replaced by the special <resource> token. Each summary in the corpora is augmented with the respective start-of-summary <start> and end-of-summary <end> tokens. The former acts a signal with which the decoder initialises the text generation process whereas the latter is outputted when the summary has been generated completely [57,82].

4.1.2. Baselines

Due to the variety of approaches for text generation, we demonstrate the effectiveness of our system by comparing it against two baselines of different nature.

Machine translation (MT) For the MT baseline, we used Google Translate on English Wikipedia summaries. Those translations are compared to the actual target language’s Wikipedia entry. This limits us to articles that exist in both English and the target language. In our dataset, the concepts in Esperanto and Arabic that are not covered by English Wikipedia account for 4.3% and 30.5% respectively. This indicates the content coverage gap between different Wikipedia languages [31].

Template retrieval (TP) Similar to template-based approaches for text generation [18,75], we build a template-based baseline that retrieves an output summary from the training data based on the input triples. First, the baseline encodes the list of input triples that corresponds to each summary in the training/test sets into a sparse vector of TF-IDF weights [35]. Afterwards, it performs LSA [28] to reduce the dimensionality of that vector. Finally, for each item in the test set, we employ the K-nearest neighbors algorithm to retrieve the vector from the training set that is

the closest to this item. The summary that corresponds to the retrieved vector is used as the output summary for this item in the test set. We provide two versions of this baseline. The first one (TP) retrieves the raw summaries from the training dataset. The second one (TP_{ext}) retrieves summaries with the special tokens for vocabulary extension. A summary can act as a template after replacing its entities with their corresponding *Property Placeholders* (see Table 1).

KN The KN baseline is a 5-gram Kneser–Ney (KN) [30] language model. KN has been used before as a baseline for text generation from structured data [51] and provided competitive results on a single domain in English. We also introduce a second KN model (KN_{ext}), which is trained on summaries with the special tokens for property placeholder. During test time, we use beam search of size 10 to sample from the learned language model.

4.2. RQ1 – Judgement-based evaluation

We defined quality in terms of text fluency and appropriateness, where fluency refers to how understandable and grammatically correct a text is, and appropriateness captures how well the text ‘feels like’ Wikipedia content. We asked two sets of participants from two different language Wikipedias to assess the same summary sentences on a scale according to these two metrics.

Participants were asked to fill out a survey combining fluency and appropriateness questions. An example for a question can be found in Fig. 5.

Question 69.

من فضلك قم بتقييم جودة النص على مقياس من صفر 0 إلى ستة 6

خماسي كلوريد الزرنيخ مركب كيميائي له الصيغة (كلمة ناقصة) ، ويكون على شكل بلورات بيضاء.

قم بتقييم جودة هذا النص:

0
 1
 2
 3
 4
 5
 6

Question 70.

قيم إذا كنت تعتقد أن هذا النص من الممكن ان يكون مقتبس من ا لويكيبيديا العربية ام لا.

لا تعتمد على أي مصادر خارجية لمعرفة الإجابة (مثل محرك بحث جوجل أو ويكيبيديا)

خماسي كلوريد الزرنيخ مركب كيميائي له الصيغة (كلمة ناقصة) ، ويكون على شكل بلورات بيضاء.

هل تعتقد أن الجملة السابقة بإمكانها أن تستخدم كأول جملة في مقال من مقالات ويكيبيديا العربية ؟

Fig. 5. Example of a question to the editors about quality and appropriateness for Wikipedia of the generated summaries in Arabic. They see this page after the instructions are displayed. First, the user is asked to evaluate the quality from 0 to 6 (Question 69), then they are asked whether the sentence could be part of Wikipedia (Question 70). Translation (Question 69): Please evaluate the text quality (0–6). (The sentence to evaluate has a grey background.) How well written is this sentence? Translation (Question 69): Please evaluate whether you think this could be a sentence from Wikipedia. Do not use any external tools (e.g. Google or Wikipedia) to answer this question. (The sentence to evaluate has a grey background.) Could the previous sentence be part of Wikipedia?

Table 5

Judgement-based evaluation: total number of participants (*P*), total number of sentences (*S*), number of participants who evaluated at least 50% of the sentences, average and mean number of sentences evaluated per participant, and number of total annotations

		#P	#S	#P: S>50%	Avg #S/P	Median #S/P	All Ann.
Arab.	Fluency	27	60	5	15.03	5	406
	Appropriateness	27	60	5	14.78	4	399
Esper.	Fluency	27	60	3	8.7	1	235
	Appropriateness	27	60	3	8.63	1	233

4.2.1. Recruitment

Our study targets any speaker of Arabic and Esperanto who reads that particular Wikipedia, independent of their contributions to Wikipedia. We wanted to reach fluent speakers of each language who use Wikipedia and are familiar with it even if they do not edit it frequently. For Arabic, we reached out to Arabic speaking researchers from research groups working on Wikipedia-related topics. For Esperanto, as there are fewer speakers and they are harder to reach, we promoted the study on social media such as Twitter and Reddit⁶ using the researchers' accounts. The survey instructions and announcements were translated to Arabic and Esperanto.⁷ The survey was open for 15 days.

4.2.2. Participants

We recruited a total of 54 participants (see Table 5). Coincidentally, 27 of them were from each language community.

4.2.3. Ethics

The research was approved by the Ethics Committee of the University of Southampton under ERGO Number 30452.

4.2.4. Data

For both languages, we created a corpus consisting of 60 summaries of which 30 are generated through our approach, 15 are from news, and 15 from Wikipedia sentences used to train the neural network model. For news in Esperanto, we chose introductory sentences of articles in the Esperanto version of *Le Monde Diplomatique*.⁸ For news in Arabic, we did the same and used the RSS feed of BBC Arabic.⁹

4.2.5. Metrics

Each participant was asked to assess the *fluency* of 60 sentences on a scale from 0 to 6 as follows:

- (6) No grammatical flaws and the content can be understood with ease
- (5) Comprehensible and grammatically correct summary that reads a bit artificial
- (4) Comprehensible summary with minor grammatical errors
- (3) Understandable, but has grammatical issues
- (2) Barely understandable summary with significant grammatical errors
- (1) Incomprehensible summary, but a general theme can be understood
- (0) Incomprehensible summary

For each sentence, we calculated the mean fluency given by all participants and then averaging over all summaries of each category.

To assess the appropriateness, participants were asked to assess whether the displayed sentence could be part of a Wikipedia article. We tested whether a reader can tell the difference from just one sentence whether a text is appropriate for Wikipedia, using the news sentences as a baseline. This gave us an insight into whether the text produced by the neural network “feels” like Wikipedia text. Participants were asked not to use any external tools for

⁶https://www.reddit.com/r/Esperanto/comments/75rytb/help_in_a_study_using_ai_to_create_esperanto/

⁷<https://github.com/luciekaffee/Announcements>

⁸<http://eo.mondediplo.com/>, accessed on the 28th of September, 2017.

⁹<http://feeds.bbci.co.uk/arabic/middleeast/rss.xml>, accessed on the 28th of September, 2017.

this task and had to give a binary answer. Similarly to fluency, average appropriateness is calculated by averaging the corresponding scores of each summary across all annotators.

4.3. RQ2 and RQ3 – Task-based evaluation

We ran a series of semi-structured interviews with editors of six Wikipedias to get an in-depth understanding of their experience with reading and using the automatically generated text. Each interview started with general questions about the experience of the participant with Wikipedia and Wikidata, and their understanding of different aspects of these projects. The participants were then asked to open and read an ArticlePlaceholder page including text generated through the NLG algorithm as shown in Fig. 6, translated to English in Fig. 7. Finally, participants

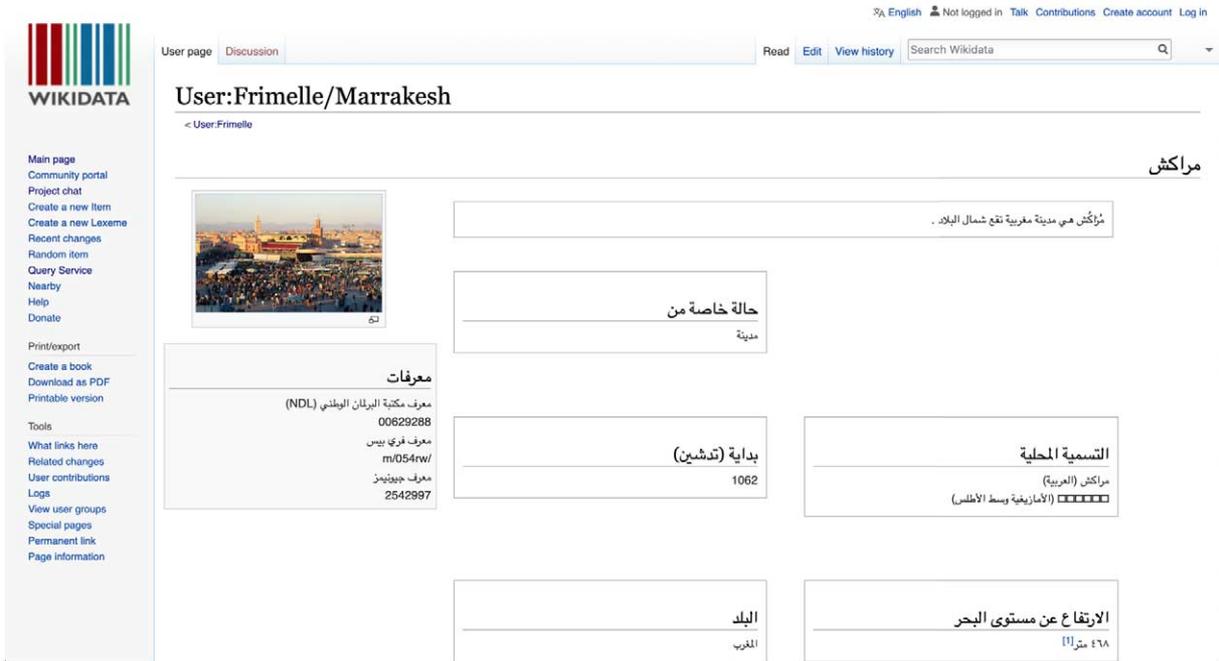


Fig. 6. Screenshot of the reading task. The page is stored as a subpage of the author’s userpage on Wikidata, therefore the layout copies the original layout of any Wikipedia. The layout of the information displayed mirrors the ArticlePlaceholder setup. The participants sees the sentence to evaluate alongside information included from the Wikidata triples (such as the image and statements) in their native language (Arabic in this example).

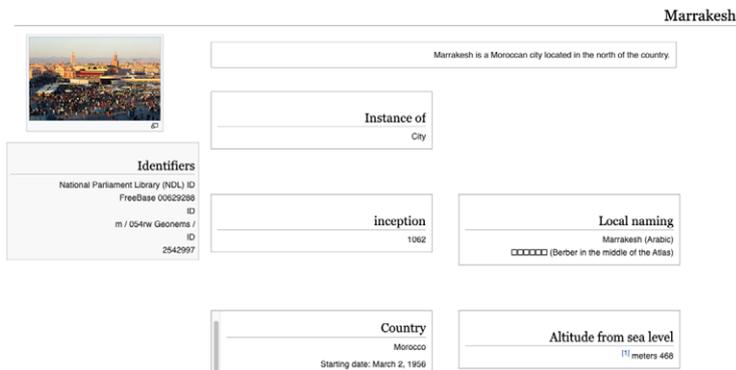


Fig. 7. Screenshot of the reading task as in Fig. 6, translated to English.

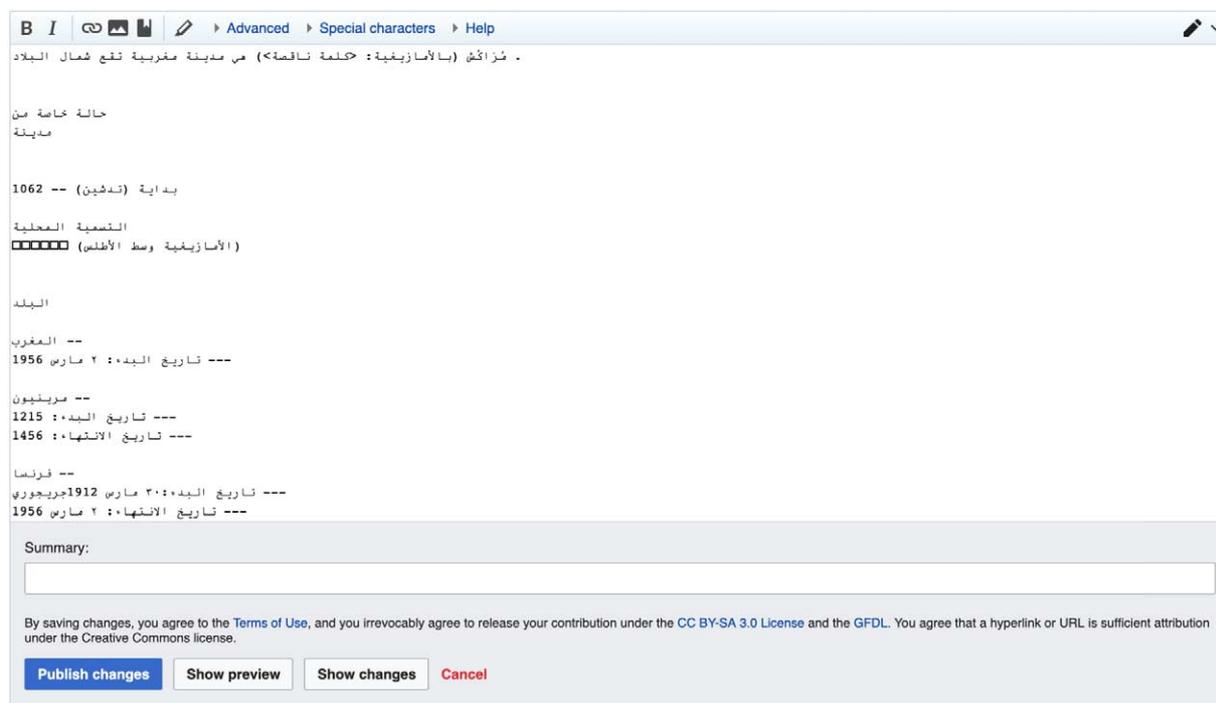


Fig. 8. Screenshot of the editing task. The page is stored on a subpage of the author's userpage on Wikidata, therefore the layout is equivalent as the current MediaWiki installations on Wikipedia. The participants see the sentence, that they saw before in the reading task and the triples from Wikidata in their native language (Arabic in this example). The triples are manually added to the page by the researchers for easier interaction with the data by the editor. The data is the same data as in the reading task (Fig. 6).

were asked to edit the content of a page, which contained the same information as the one they had to read earlier, but was displayed as plain text in the Wikipedia edit field. The editing field can be seen in Fig. 8.

4.3.1. Recruitment

The goal was to recruit a set of editors from different language backgrounds to have a comprehensive understanding of different language communities. We reached out to editors of different Wikipedia editor mailinglists¹⁰ and tweeted a call for contribution using the lead author's account.¹¹

We were in contact with 18 editors from different Wikipedia communities. We allowed all editors to participate but had to exclude editors who edit only on English Wikipedia (as it is outside our use-case) and editors who did not speak a sufficient level of English, which made conducting the interview impossible.

4.3.2. Participants

Our sample consists of 10 Wikipedia editors of different lesser resourced languages (measured in their number of articles compared to English Wikipedia). We originally conducted interviews with 11 editors from seven different language communities, but had to remove one interview with an editor of the Breton Wikipedia, as we were not able to generate the text for the reading and editing tasks because of a lack of training data.

Among the participants, 4 were from the Arabic Wikipedia and participated in the judgement-based evaluation from RQ1. The remaining 6 were from other smaller Wikipedia language communities: Persian, Indonesian, Hebrew, Ukrainian (one per language) and Swedish (two participants).

¹⁰Mailinglists contacted: wikiar-l@lists.wikimedia.org (Arabic), wikio-l@lists.wikimedia.org (Esperanto), wikifa-l@lists.wikimedia.org (Persian), Wikidata mailinglist, Wikimedia Research Mailinglist.

¹¹<https://twitter.com/frimelle/status/1031953683263750144>

Table 6

Number of Wikipedia articles, active editors on Wikipedia (editors that performed at least one edit in the last 30 days), and number of native speakers in million

Language	# Articles	Active editors	Native speakers
Arabic	541,166	5,398	280
Swedish	3,763,584	2,669	8.7
Hebrew	231,815	2,752	8
Persian	643,635	4,409	45
Indonesian	440,948	2,462	42.8
Ukrainian	830,941	2,755	30

Table 7

Overview of sentences and number of participants in each language

Language	Sentence displayed to participants	# Participants	Participant
Arabic (ar)	مُرَّاكُش (بالأمازيغية: <rare>) هي مدينة مغربية تقع شمال البلاد.	4	P1, P4, P8, P9
Swedish (sv)	Marrakech (arabiska <rare>, Berberspråk <rare>) är en stad i sydvästra Marocko, vid foten av <rare>.	2	P5, P6
Hebrew (he)	מרקש (בערבית: <rare>) היא עיר מדברית בדרום מערב מרוקו למרגלות הרי <rare>.	1	P3
Persian (fa)	شهر مراکش (به بربری: <rare>) یکی از شهرهای کشور مراکش و مرکز استان مراکش <rare> است.	1	P2
Indonesian (id)	Marrakesh (Arab: <rare>) ialah kota di barat daya Maroko di kaki <rare>.	1	P7
Ukrainian (uk)	Марракеш (араб. <rare>) — важливе імперське місто в Марокко, розташованого біля підніжжя гір <rare>.	1	P10

While Swedish is officially the third largest Wikipedia in terms of number of articles,¹² most of the articles are not manually edited. In 2013, the Swedish Wikipedia passed one million articles, thanks to a bot called *lsjbot*, which at that point in time had created almost half of the articles on Swedish Wikipedia.¹³ Such bot-generated articles are commonly limited, both in terms of information content and length. The high activity of a single bot is also reflected in the small number of active editors compared to the large number of articles (see Table 6).

The participants were experienced Wikipedia editors, with average tenures of 9.3 years in Wikipedia (between 3 and 14 years, median 10). All of them have contributed to at least one language besides their main language, and to the English Wikipedia. Further, 4 editors worked in at least two other languages beside their main language, while 2 editors were active in as many as 4 other languages. All participants knew about Wikidata, but had varying levels of experience with the project. 4 participants have been active on Wikidata for over 3 years (with 2 editors being involved since the start of the project in 2013), 5 editors had some experience with editing Wikidata and one editor had never edited Wikidata, but knew the project.

Table 7 displays the sentence used for the interviews in different languages. The Arabic sentence is generated by the network based on Wikidata triples, while the other sentences are synthetically created as described below.

¹²https://meta.wikimedia.org/wiki/List_of_Wikipedias

¹³<https://blog.wikimedia.org/2013/06/17/swedish-wikipedia-1-million-articles/>

4.3.3. Ethics

The research was approved by the Ethics Committee of the University of Southampton under ERGO Number 44971 and written consent was obtained from each participant ahead of the interviews.

4.3.4. Data

For Arabic, we reused a summary sentence from the *RQ1* evaluation. For the other language communities, we emulated the sentences the network would produce. First, we picked an entity to test with the editors. Then, we looked at the output produced by the network for the same concept in Arabic and Esperanto to understand possible mistakes and tokens produced by the network. We chose the city of *Marrakesh* as the concept editors would work on. Marrakesh is a good starting point, as it is a topic possibly highly relevant to readers and is widely covered, but falls into the category of topics that are potentially under-represented in Wikipedia due to its geographic location [24]. An article about Marrakesh exists in 93 language editions (as of September 2020), including the ones of this study.

We collected the introductory sentences for Marrakesh in the editors' languages from Wikipedia. Those are the sentences the network would be trained on and tries to reproduce. We ran the keyword matcher that was used for the preparation of the dataset on the original Wikipedia sentences. It marked the words the network would pick up or would be replaced by property placeholders. Therefore, these words could not be removed.

As we were particularly interested in how editors would interact with the missing word tokens, the network can produce, we removed up to two words in each sentence: the word for the concept in its native language (e.g. *Morocco* in Arabic for non-Arabic sentences), as we saw that the network does the same, and the word for a concept that is not connected to the main entity of the sentence on Wikidata (e.g. *Atlas Mountains*, which is not linked to Marrakesh). An overview of all sentences used in the interviews can be found in Table 7.

4.3.5. Task

The interview started with an opening, explaining that the researcher will observe the reading and editing of the participant in their language Wikipedia. Until both reading and editing were finished, the participant did not know about the provenance of the text. To start the interview, we asked demographic questions about the participants' contributions to Wikipedia and Wikidata, and to test their knowledge on the existing ArticlePlaceholder. Before reading, they were introduced to the idea of displaying content from Wikidata on Wikipedia. Then, they saw the mocked page of the ArticlePlaceholder as can be seen in Fig. 6 in Arabic. Each participant saw the page in their respective language. As the interviews were remote, the interviewer asked them to share the screen so they could point out details with the mouse cursor. Questions were asked to let them describe their impression of the page while they were looking at the page. Then, they were asked to open a new page, which can be seen in Fig. 8. Again, this page would contain information in their language. They were asked to edit the page and describe what they are doing at the same time freely. We asked them to not edit a whole page but only write two to three sentences as the introduction to a Wikipedia article on the topic with as much of the information given as needed. After the editing was finished, they were asked questions about their experience. (For the interview guideline, see Appendix A.) The interview followed the methodology of a semi-structured interview in which all participants were asked the same questions. Only then, the provenance of the sentences was revealed. Given this new information, we asked them about the predicted impact on their editing experience. Finally, we left them time to discuss open questions of the participants. The interviews were scheduled to last between 20 minutes to one hour.

4.3.6. Analysing the interviews

We interviewed a total of 11 editors of seven different language Wikipedias. The interviews took place in September 2018. We used thematic analysis to evaluate the results of the interviews. The interviews were coded by two researchers independently, in the form of inductive coding based on the research questions. After comparing and merging all themes, both researchers independently applied these common themes on the text again.

4.3.7. Editors' reuse metric

Editors were asked to complete a writing task. We assessed how they used the automatically generated summary sentences in their work by measuring the amount of text reuse. We based the assessment on the editors' resultant summaries after the interviews were finished.

To quantify the amount of reuse in text we use the Greedy String-Tiling (GST) algorithm [89]. GST is a substring matching algorithm that computes the degree of reuse or copy from a source text and a dependent one. GST is able to deal with cases when a whole block is transposed, unlike other algorithms such as the Levenshtein distance, which calculates it as a sequence of single insertions or deletions rather than a single block move. Adler and de Alfaro [2] introduce the concept of *Edit Distance* in the context of vandalism detection on Wikipedia. They measure the trustworthiness of a piece of text by measuring how much it has been changed over time. However, their algorithm punishes the copy of the text, as they measure every edit to the original text. In comparison, we want to measure how much of the text is reused, therefore GST is appropriate for the task at hand.

Given a generated summary $S = s_1, s_2, \dots$ and an edited one $D = d_1, d_2, \dots$, each consisting of a sequence of tokens, GST identifies a set of disjoint longest sequences of tokens in the edited text that exist in the source text (called *tiles*) $T = \{t_1, t_2, \dots\}$. It is expected that there will be common stop words appearing in both the source and the edited text. However, we were rather interested in knowing how much of the real structure of the generated summary is being copied. Thus, we set minimum match length factor $mml = 3$ when calculating the tiles, s.t. $\forall t_i \in T : t_i \subseteq S \wedge t_i \subseteq D \wedge |t_i| \geq mml$ and $\forall t_i, t_j \in T | i \neq j : t_i \cap t_j = \emptyset$. This means that copied sequences of single or double words will not count in the calculation of reuse. We calculated a reuse score $gstscore$ by counting the lengths of the detected tiles, and normalized by the length of the generated summary.

$$gstscore(S, D) = \frac{\sum_{t_i \in T} |t_i|}{|S|} \quad (1)$$

We classified each of the edits into three groups according to the $gstscore$ as proposed by [13]: 1) **Wholly Derived (WD)**: the summary sentence has been fully reused in the composition of the editor’s text ($gstscore \geq 0.66$); 2) **Partially Derived (PD)**: the summary sentence has been partially used ($0.66 > gstscore \geq 0.33$); 3) **Non Derived (ND)**: the text has been changed completely ($0.33 > gstscore$).

5. Results

5.1. RQ1 – Metric-based corpus evaluation

As noted earlier, the evaluation used standard metrics in this space and data from the Arabic and Esperanto Wikipedias. We compared against five baselines, one in machine translation (*MT*), two based on templates (*TP* and *TP_{ext}*), the other main class of techniques in NLG, and two language models (*KN* and *KN_{ext}*). As displayed in Table 8, our model showed a significant enhancement compared to all baselines across the majority of the evaluation metrics in both languages. We achieved a 3.01 and 5.11 enhancement in BLEU 4 score in Arabic and Esperanto respectively over *TP_{ext}*, the strongest baseline.

The tests also hinted at the limitations of using machine translation for this task. We attributed this result to the different writing styles across language versions of Wikipedia. The data confirms that generating language from labels of conceptual structures such as Wikidata is a much more suitable approach.

Around $\leq 1\%$ of the generated summary sentences in Arabic and 2% of the Esperanto ones did not end in `<end>` token and were hence incomplete. We assumed the difference between the two was due to the size of the Esperanto dataset, which makes it harder for the trained models (i.e., with or without property placeholders) to generalise on unseen data (i.e. summaries for which the special end-of-summary `<end>` token is generated). These snippets were excluded from the evaluation.

The introduction of the property placeholders to our encoder-decoder architecture enhances our performance further by 0.61–1.10 BLEU (using BLEU 4).

In general, our property placeholder mechanism benefits the performance of all the competitive systems.

Generalisation across domains Summaries related to years, galaxies, settlements and villages in Esperanto were found to be among the top performing domains, consistently achieving a BLEU-4 performance greater than 50. In Arabic, there are more than 13 different domains on which our system sustains an average BLEU-4 score greater

Table 8

Automatic evaluation of our model against all other baselines using BLEU 2–4, ROUGE and METEOR for both Arabic and Esperanto validation and test set

	Model	BLEU 2		BLEU 3		BLEU 4		ROUGE _L		METEOR	
		Valid.	Test	Valid.	Test	Valid.	Test	Valid.	Test	Valid.	Test
Arabic	KN	2.28	2.4	0.95	1.04	0.54	0.61	17.08	17.09	29.04	29.02
	KN _{ext}	21.21	21.16	16.78	16.76	13.42	13.42	28.57	28.52	30.47	30.43
	MT	19.31	21.12	12.69	13.89	8.49	9.11	31.05	30.1	29.96	30.51
	TP	34.18	34.58	29.36	29.72	25.68	25.98	43.26	43.58	32.99	33.33
	TP _{ext}	42.44	41.5	37.29	36.41	33.27	32.51	51.66	50.57	34.39	34.25
	Ours	47.38	48.05	42.65	43.32	38.52	39.20	64.27	64.64	45.89	45.99
	w/ PrP	47.96	48.27	43.27	43.60	39.17	39.51	64.60	64.69	46.09	46.17
Esperanto	KN	6.91	6.64	4.18	4.0	2.9	2.79	37.48	36.9	31.05	30.74
	KN _{ext}	16.44	16.3	11.99	11.92	8.77	8.79	44.93	44.77	33.77	33.71
	MT	1.62	1.62	0.59	0.56	0.26	0.23	0.66	0.68	4.67	4.79
	TP	33.67	33.46	28.16	28.07	24.35	24.3	46.75	45.92	20.71	20.46
	TP _{ext}	43.57	42.53	37.53	36.54	33.35	32.41	58.15	57.62	31.21	31.04
	Ours	42.83	42.95	38.28	38.45	34.66	34.85	66.43	67.02	40.62	41.13
	w/ PrP	43.57	43.19	38.93	38.62	35.27	34.95	66.73	66.61	40.80	40.74

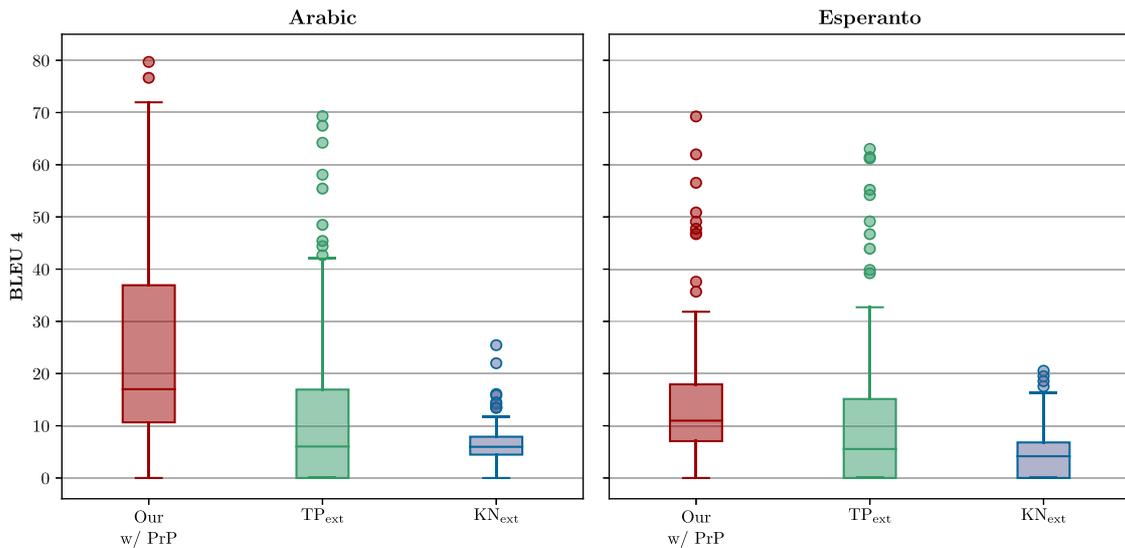


Fig. 9. A box plot showing the distribution of BLEU 4 scores of all systems for each category of generated summaries.

than 50; GrandPrix, fern, crustacean, village and cycad are among the top performing domains, whose summaries are scored with BLEU-4 scores of 66 or higher. Volcano, food and holiday, and anatomical structure, magazine and ethnic group are the three domains on which our system achieves the lowest BLEU-4 performance in Esperanto and Arabic respectively.

To investigate how well different models can generalise across multiple domains, we categorise each generated summary into one of 50 categories according to its main entity instance type (e.g. village, company, football player). We examine the distribution of BLEU-4 scores per category to measure how well the model generalises across domains (Fig. 9).

We show that i) the high performance of our system is not skewed towards some domains at the expense of others, and that ii) our model has a good generalisation across domains – better than any other baseline.

For instance, the over performance of TP_{ext} is limited to a small number of domains – plotted as the few outliers in Fig. 9 for TP_{ext} –, despite its performance being much lower on average for all the domains.

Despite the fact that TP_{ext} achieves the highest recorded performance in a few domains (i.e. TP_{ext} outliers in Fig. 9), its performance is much lower on average for all the domains. The valuable generalisation of our model across domains is mainly due to the language model in the decoder layer of our model, which is more flexible than rigid templates and can adapt easier to multiple domains. Despite the fact that the Kneser–Ney template-based baseline (KN_{ext}) has exhibited competitive performance in a single-domain context [51], it is failing to generalise in our multi-domain text generation scenario. Unlike our approach, KN_{ext} does not incorporate the input triples directly for generating the output summary, but rather only uses them to replace the special tokens after a summary has been generated. This might yield acceptable performance in a single domain, where most of the summaries share a very similar pattern. However, it struggles to generate a different pattern for each input set of triples in multiple domain summary generation.

5.2. RQ1 – Judgement-based evaluation

5.2.1. Fluency

As shown in Table 9, overall, the quality of the generated text is high (4.7 points out of 6 in average in Arabic and 4.5 in Esperanto). In Arabic, 63.3% of the summaries scored as much as 5 in fluency on average. In Esperanto, which had the smaller training corpus, the participants nevertheless gave as many as half of the snippets an average of 5, with 33% of them reaching a 6 by all participants. We concluded that in most cases, the text we generated was very understandable and grammatically correct. In addition, the results were perceived to match the quality of writing in Wikipedia and news reporting.

5.2.2. Appropriateness

The results for appropriateness are summarised in Table 9. A majority of the snippets were considered to be part of Wikipedia (77% for Arabic and 69% in Esperanto). The participants confirmed that they seemed to identify a certain style and manner of writing with Wikipedia – by comparison, only 35% of the Arabic news snippets and 52% of the Esperanto ones could have passed as Wikipedia content in the study. Genuine Wikipedia text was recognised as such (in 77% and 84% of the cases, respectively).

Our model was able to generate text that is not only accurate from a writing point of view, but in a high number of cases, felt like Wikipedia and could blend in with other Wikipedia content.

5.3. RQ2 task-based evaluation

As part of our interview study, we asked editors to read an ArticlePlaceholder page with included NLG text and asked them to comment on a series of issues. We grouped their answers into several general themes around: their use of the snippets, their opinions on text provenance, the ideal length of the text, the importance of the text for the ArticlePlaceholder, and limitations of the algorithm.

Table 9
Results for fluency and appropriateness

		Fluency		Appropriateness
		Mean	SD	Part of Wikipedia
Arabic	Ours	4.7	1.2	77% 
	Wikipedia	4.6	0.9	74% 
	News	5.3	0.4	35% 
Esper.	Ours	4.5	1.5	69% 
	Wikipedia	4.9	1.2	84% 
	News	4.2	1.2	52% 

Use The participants appreciated the summary sentences:

“I think it would be a great opportunity for general Wikipedia readers to help improve their experience, while reading Wikipedia” (P7).

Some of them noted that the summary sentence on the ArticlePlaceholder page gave them a useful overview and quick introduction to the topic of the page, particularly for people trained in one language or non-English speakers:

“I think that if I saw such an article in Ukrainian, I would probably then go to English anyway, because I know English, but I think it would be a huge help for those who don’t” (P10).

Provenance As part of the reading task, we asked the editors what they believed was the provenance of the information displayed on the page. This gave us more context to the promising fluency and appropriateness scores achieved in the quantitative study. The editors made general comments about the page and tended to assume that the generated sentence was from other Wikipedia language versions:

[The generated sentence was] “taken from Wikipedia, from Wikipedia projects in different languages.” (P1)

Editors more familiar with Wikidata suggested the information might be derived from Wikidata’s descriptions:

“it should be possible to be imported from Wikidata” (P9).

Only one editor could spot a difference in the generated sentence (text) and regular Wikidata triples:

“I think it’s taken from the outside sources, the text, the first text here, anything else I don’t think it has been taken from anywhere else, as far as I can tell” (P2).

Overall, the answers supported our assumption that NLG, trained on Wikidata labelled triples, could be naturally added to Wikipedia pages without changing the reading experience. In the same time, the task revealed questions around algorithmic complexity and capturing provenance. Both are relevant to ensure transparency and accountability and help flag quality issues.

Length We were interested in understanding how we could iterate over the NLG capabilities of our system to produce text of appropriate length. While the model generated just one sentence, the editors thought it to be a helpful starting point:

“Actually I would feel pretty good learning the basics. What I saw is the basic information of the city so it will be fine, almost like a stub” (P4).

While generating larger pieces of text could arguably be more useful, reducing the need for manual editing even further, the fact that the placeholder page contained just one sentence made it clear to the editors that the page still requires work. In this context, another editor referred to a ‘*magic threshold*’ for an automatically generated text to be useful (see also Section 5.4). Their expectations for an NLG output were clear:

“So the definition has to be concise, a little bit not very long, very complex, to understand the topic, is it the right topic you’re looking for or”. (P1)

We noted that whatever the length of the snippet, it needs to match reading practice. Editors tend to skim articles rather than reading them in detail:

“[...] most of the time I don’t read the whole article, it’s just some specific, for instance a news piece or some detail about I don’t know, a program in languages or something like that and after that, I just try to do something with the knowledge that I learned, in order for me to acquire it and remember it” (P6) “I’m getting more and more convinced that I just skim” (P1) “I should also mention that very often, I don’t read the whole article and very often I just search for a particular fact” (P3) “I can’t say that I read a lot or reading articles from the beginning to the end, mostly it’s getting, reading through the topic, “Oh, what this weird word means,” or something” (P10)

When engaging with content, people commonly go straight to the part of the page that contains what they need. If they are after an introduction to a topic, having a summary at the top of the page, for example in the form of an automatically generated summary sentence, could make a real difference in matching their information needs.

Importance When reading the ArticlePlaceholder page, people looked first at our text:

“The introduction of this line, that’s the first thing I see” (P4).

This is their way to confirm that they landed on the right page and if the topic matches what they were looking for:

“Yeah, it does help because that’s how you know if you’re on the right article and not a synonym or some other article” (P1).

This makes the text critical for the engagement with the ArticlePlaceholder page, where most of the information is expressed as Wikidata triples. Natural language can add context to structured data representations:

“Well that first line was, it’s really important because I would say that it [the ArticlePlaceholder page] doesn’t really make a lot of sense [without it] . . . it’s just like a soup of words, like there should be one string of words next to each other so this all ties in the thing together. This is the most important thing I would say.” (P8)

<rare> tags To understand how people react to a fault in the NLG algorithm, we chose to leave the *<rare>* tags in the summary sentences the participants saw during the reading task. As mentioned earlier, such tags refer to entities in the triples that the algorithm was unsure about and could not verbalise. We did not explain the meaning of the tokens to the participants beforehand and none of the editors mentioned them during the interviews. We believe this was mainly because they were not familiar with what the tokens meant and not because they were not able to spot errors overall. For example, in the case of Arabic, the participants pointed to a Wikidata property with an incorrect label, which our NLG algorithm reused. They also picked up on a missing label in the native language for a city. However, the *<rare>* tokens were not noticed in any of the 10 reading tasks until explicitly mentioned by the interviewer. The name of the city Marrakesh in one of the native languages (Berber) was realised using the *<rare>* token (see the Arabic sentence in Table 7). One editor explained that the fact that they are not familiar with this language (Berber) and can therefore not evaluate the correctness of the statement is the main reason that they oversaw the token:

“the language is specific, it says that this is a language that is spoken mainly in Morocco and Algeria, the language, I don’t even know the symbols for the alphabets [. . .] I don’t know if this is correct, I don’t know the language itself so for me, it will go unnoticed. But if somebody from that area who knows anything about this language, I think they might think twice” (P8).

5.4. RQ3 – Task-based evaluation

Our third research questions focused on how people work with automatically generated text. The overall aim of adding NLG to the ArticlePlaceholder is to help Wikipedia editors bootstrap missing articles without disrupting their editing practices.

As noted earlier, we carried out a task-based evaluation, in which the participants were presented with an ArticlePlaceholder page that included the summary sentence and triples from Wikidata relevant to the topic. We carried out a quantitative analysis of the editing activities via the GST score, as well as a qualitative, thematic analysis of the interviews, in which the participants explained how they changed the text. In the following we will first present the GST scores, and then discuss the themes that emerged from the interviews.

Reusing the text As shown in Table 10, the snippets are heavily used and all participants reused them at least partially. 8 of them were wholly derived (*WD*) and the other 2 were partially derived (*PD*) from the text we provided, which means an average GST score of 0.77. These results are in line with a previous editing study of ours, described in [37], with a sample of 54 editors from two language communities, 79% and 93% of the snippets were either wholly (*WD*) or partially (*PD*) reused.

Table 10

Number of snippets in each category of reuse. A generated snippet (top) and its edited version (bottom). Solid lines represent reused tiles, while dashed lines represent overlapping sub-sequences not contributing to the *gstscore*. The first two examples are created in the studies, the last one (ND) is from a previous experiment. a list of all created sentences and their translations to English can be found in Appendix B

Category	Examples	#
WD	Marrakesh (arab. <rare>) — важливе імперське місто в Марокко, розташованого біля підніжжя гір <rare>. <small>A</small> ----- <small>B</small>	8
	Marrakesh (arab. مراکش) — важливе імперське місто в Марокко, розташованого біля підніжжя гір. <small>A</small> ----- <small>B</small>	
PD	Marrakech (arabiska <rare>, Berberspråk <rare>) är en stad i sydvästra Marocko, vid foten av <rare>. <small>C</small> ----- <small>D</small> ----- <small>E</small> -----	2
	Marrakech (arabiska: مراکش, tamazight: ⵎⴰⵔⴰⴽⵏ) är en stad i Marocko med 928 850 invånare (2014). <small>C</small> ----- <small>D</small> ----- <small>E</small> -----	
ND	مُرَّاكش (بالأمازيغية: <rare>) هي مدينة مغربية تقع شمال البلاد. <small>F</small> -----	0

We manually inspected all edits and compared them to the ‘originals’ – as explained in Section 4, we had 10 participants from 6 language communities, who edited 6 language versions of the same article. In the 8 cases where editors reused more of the text, they tended to copy it with minimal modifications, as illustrated in sequences *A* and *B* in Table 10. Three editors did not change the summary sentence at all (including the special token), but only added to it based on the triples shown on the ArticlePlaceholder page.

One of the common things that hampers the full reusability are <rare> tokens. This can lead editors to rewrite the sentence completely, as in the PD example in Table 10.

Editing experience While GST scores gave us an idea about the extent to which the automatically generated text is reused by the participants, the interviews helped us understand how they experienced the text. Overall, the summary sentences were widely praised as helpful, especially for newcomers:

“Especially for new editors, they can be a good starting help: “I think it would be good at least to make it easier for me as a reader to build on the page if I’m a new volunteer or a first time edit, it would make adding to the entry more appealing (...) I think adding a new article is a barrier. For me I started only very few articles, I always built on other contribution. So I think adding a new page is a barrier that Wikidata can remove. I think that would be the main difference.” (P4)

All participants were experienced editors. Just like in the reading task, they thought having a shorter text to start editing had advantages:

“It wasn’t too distracting because this was so short. If it was longer, it would be (...) There is this magical threshold up to which you think that it would be easier to write from scratch, it wasn’t here there.” (P10)

The length of the text is also related to the ability of the editors to work around and fix errors, such as the <rare> tokens discussed earlier. When editing, participants were able to grasp what information was missing and revise the text accordingly:

“So I have this first sentence, which is a pretty good sentence and then this is missing in Hebrew and well, since it’s missing and I do have this name here, I guess I could quickly copy it here so now it’s not missing any longer.” (P3)

The same participant checked the Wikidata triples listed on the same ArticlePlaceholder page to find the missing information, which was not available there either, and then looked it up on a different language version of Wikipedia. They commented:

“This first sentence at the top, was it was written, it was great except the pieces of information were missing, I could quite easily find them, I opened the different Wikipedia article and I pasted them, that was really nice” (P3).

Other participants mentioned a similar approach, though some decided to delete the entire snippet because of the `<rare>` token and start from scratch. However, the text they added turned out to be very close to what the algorithm generated.

“I know it, I have it here, I have the name in Arabic, so I can just copy and paste it here.” (P10)

“[deleted the whole sentence] mainly because of the missing tokens, otherwise it would have been fine” (P5)

One participant commented at length on the presence of the tokens:

“I didn’t know what rare [is], I thought it was some kind of tag used in machine learning because I’ve seen other tags before but it didn’t make any sense because I didn’t know what use I had, how can I use it, what’s the use of it and I would say it would be distracting, if it’s like in other parts of the page here. So that would require you to understand first what rare does there, what is it for and that would take away the interest I guess, or the attention span so it would be better just to, I don’t know if it’s for instance, if the word is not, it’s rare, this part right here which is, it shouldn’t be there, it should be more, it would be better if it’s like the input box or something”. (P1)

Overall, the editing task and the follow-up interviews showed that the summary sentences were a useful starting point for editing the page. Missing information, presented in the form of `<rare>` markup did not hinder participants from editing and did not make them consider the snippets less useful. While they were unsure about what the tokens meant, they intuitively replaced them with the information they felt was missing, either by consulting the Wikidata triples that had not been summarised in the text, or by trying to find that information elsewhere on Wikipedia.

6. Discussion

Our first research question focuses on how well an NLG algorithm can generate summaries from the Wikipedia reader’s perspective. In most of the cases, the text is considered to be from the Wikimedia environment. Readers do not clearly differentiate between the generated summary sentence and an original Wikipedia sentence. While this indicates the high quality of the generated textual content, it is problematic with respect to trust in Wikipedia. Trust in Wikipedia and how humans evaluate trustworthiness of a certain article has been investigated using both quantitative and qualitative methods. Adler et al. [1] and Adler and de Alfaro [2] develop a quantitative framework based on Wikipedia’s history. Lucassen and Schraagen [56] use a qualitative methodology to code readers’ opinions on the aspects that indicate the trustworthiness of an article. However, none of these approaches take the automatic creation of text by non-human algorithms into account. A high quality Wikipedia summary, which is not distinguishable from a human-generated one, can be a double-edged sword. While conducting the interviews, we realized that the Arabic generated summary has a factual mistake. We could show in previous work [84] that those factual mistakes are relatively seldom, however they are a known drawback of neural text generation. In our case, the Arabic sentence stated that Marrakesh was located in the north, while it is actually in the center of the country. One of the participants lives in Marrakesh. Curiously, they did not realize this mistake, even while translating the sentence to the interviewee:

“Yes, so I think, so we have here country, Moroccan city in the north, I would say established and in year (...)” (P1)

As we did not introduce the sentence as automatically generated, we assume it is due to the trust Wikipedians have in their platform and the quality of the sentence:

“This sentence was so well written that I didn’t even bother to verify if it’s actually a desert city” (P3)

To not misinform and eventually disrupt this trust, future work will have to investigate ways of dealing with such unfactual statements. On a technical level that will mean exploring ways to excluding sentences with factual mistakes, called hallucinations. Those hallucinations can be found across different models, including the most recent self-attentive NLG architectures [44]. On an HCI level, investigating ways of communicating the problems of neural text generation to the reader will be needed. One possible solution to this problem can be visualizations, as these

can influence the trust given to a particular article [43], such as WikiDashboard [68]. A similar tool could be used for the provenance of text.

Supporting the previous results from the readers, editors have also a positive perception of the summaries. It is the first thing they read when they arrive at a page and it helps them to quickly verify that the page is about the topic they are looking for.

When creating the summaries, we assumed their relatively short length might be a point for improvement from an editors' perspective. In particular, as research suggests that the length of an article indicates its quality – basically the longer, the better [9]. From the interviews with editors, we found that they mostly skim articles when reading them. This seems to be the more natural way of browsing the information on an article and is supported by the short summary, giving an overview on the topic.

All editors we worked with are part of the multilingual Wikipedia community, editing in at least two Wikipedias. Hale [26,27] highlight that users of this community are particularly active compared to their monolingual counterparts and confident in editing across different languages. However, taking potential newcomers into account, they suggest that the ArticlePlaceholder might be helpful to lower the barrier of starting to edit. Recruiting more editors has been a long-standing objective, with initiatives such as the Tea House [62] aiming at welcoming and comforting new editors; Wikipedia Adventure employs a similar approach using a tool with gamification features [63]. The ArticlePlaceholder, and in particular the provided summaries in natural language, can have an impact on how people start editing.

In comparison to Wikipedia Adventure, the readers are exposed to the ArticlePlaceholder pages and, thus, it could lower their reservation to edit by offering a more natural start of editing.

Lastly, we asked the research question how editors use the textual summaries in their workflow. Generally, we can show that the text is highly reused. One of the editors mentions a *magic threshold*, that makes the summary acceptable as a starting point for editing. This seems similar to post-editing in machine translation (or rather monolingual machine translation [33], where a user only speaks the target or source language). Translators have been found to oppose machine translation and post-editing as they perceive it as more time consuming and restricting with respect to their freedom in the translation (e.g. sentence structure) [49]. Nonetheless, it has been shown that a different interface can not only lead to reduced time and enhanced quality, but also convinces a user to believe in the improved quality of the machine translation [25]. This underlines the importance of the right integration with machine-generated sentences, as we aim for in the ArticlePlaceholder.

The core of this work is to understand the perception of a community, such as Wikipedians, of the integration of a state-of-the-art machine learning technique for NLG in their platform. We can show that an integration can work and be supported. This finding aligns with other projects already deployed on Wikipedia. For instance, bots (short for robots) that monitor Wikipedia, have become a trusted tool for vandalism fighting [23]; so much that they can even revert edits made by humans if they believe them to be malicious. The cooperative work between humans and machines on Wikipedia has been also theorized in machine translation. Alegria et al. [3] argue for the integration of machine translation in Wikipedia, that learns from the post-editing of the editors. Such a human-in-the-loop approach is also applicable to our NLG work, where a algorithm could learn from the humans' contributions.

There is a need of investigating this direction further, as NLG algorithms will not achieve the same quality as humans. Especially in a low resource setting, as the one observed in this work, human support is needed. However, automated tools can be a great way of allocating the limited human resources to the tasks that are mostly needed. Post-editing the summaries can serve a purely data-driven approach such as ours with additional data that can be used to further improve the quality of the automatically generated content. To make such an approach feasible for the ArticlePlaceholder, we need an interface that encourages the editing of the summaries. The less effort this editing requires, the more we can ensure an easy collaboration of human and machine.

7. Limitations

We interviewed ten editors having different levels of Wikipedia experience. As all editors are already Wikipedia editors, the conclusions we can draw for new editors are limited. We focus on experienced editors, as we expect

them to be the first editors to adapt the ArticlePlaceholder in their workflow. Typically, new editors will follow the existing guidelines and standards of the experienced editors, therefore, the focus on experienced editors will give us an understanding of how the editors will accept and interact with the new tool. Further, it is difficult to sample from new editors, as there is a variety of factors that can make a contributor develop into a long-term editor or not [48].

The distribution of languages favours Arabic, as the community was most responsive. This can be assumed to be due to previous collaborations. While we cover different languages, it is only a small part of the different language communities that Wikipedia covers in total. Most studies of Wikipedia editors currently focus on English Wikipedia [65]. Even the few studies that observe multiple language Wikipedia editors do not include the span of insights from different languages that we provide in this study. In our study we treat the different editors as members of a unified community of Wikipedia underserved languages. This is supported by the fact that their answers and themes were consistent across different languages. Therefore, adding more editors of the same languages would not have brought a benefit.

We aimed our evaluation at two populations: readers and editors. While the main focus was on the editors and their interaction with the new information, we wanted to include the readers' perspective. In the readers evaluation we focus on the quality of text (in terms of fluency and appropriateness), as this will be the most influential factor for their experience on Wikipedia. Readers, while usually overseen, are an important part of the Wikipedia community [52]. Together, those two groups form the Wikipedia community as new editors are recruited from the existing pool of readers and readers contribute in essential ways to Wikipedia as shown by Antin and Cheshire [5].

The sentences the editors worked on are synthesized, i.e. not automatically generated but created by the authors of this study. An exception is the Arabic sentence, which was generated by the approach described in Section 4. While those synthesized sentences were not created by natural language generation, they were created and discussed with other researchers in the field. Our goal was to explore the usability and limitations of recent data-driven approaches w.r.t. Wikipedia's editing community and the extent to which such approaches can effectively support their work. Limitations that currently prohibit the wide usability of these approaches are mostly associated with hallucinations and manifestation of rare tokens, and remain relevant to even the most recent systems. We therefore focused on the most common problem in text generative tasks similar to ours: the `<rare>` tokens. Other problems of neural language generation, such as factually wrong sentences or hallucinations [74], were excluded from the study as they are not a common problem for short summaries as ours. The extent of hallucinations on single sentence generative scenarios has also been explored by Chisholm et al. [11], who have shown a precision of 93% for generation of English biographies. However, we believe this topic should be explored further in future work since as we show in our study factual mistakes caused by hallucinations can be easily missed by the editors. Further, our study set up was focusing on the integration in the ArticlePlaceholder, i.e. the quality of the sentences rather than their factual correctness. We leave it to future work to apply the mixed methods approach of evaluating natural language generation proposed in this paper to full generated articles.

8. Conclusion

We conducted a quantitative study with members of the Arabic and Esperanto Wikipedia community and semi-structured interviews with members of six different Wikipedia communities to understand the communities understanding and acceptance of generated text in Wikipedia. To understand the impact of automatically generated text for a community such as the Wikipedia editors, we surveyed their perception of generated summaries in terms of fluency and appropriateness. To deepen this understanding, we conducted 10 semi-structured interviews with experienced Wikipedia editors from 6 different language communities and measured their reuse of the original summaries.

The addition of the summaries seems to be natural for readers: We could show that Wikipedia editors rank our text close to the expected quality standards of Wikipedia, and are likely to consider the generated text as part of Wikipedia. The language the neural network produces integrates well with the existing content of Wikipedia, and readers appreciate the summary on the ArticlePlaceholder as it is the most helpful element on the page to them.

We could emphasize that editors would assume the text was part of Wikipedia and that the summary improves the ArticlePlaceholder page fundamentally. Particularly the summary being short supported the usual workflow of skimming the page for information needed. The missing word token, that we included to gain an understanding how users interact with faulty produced text, did not hinder the reading experience nor the editing experience. Editors are likely to reuse a large portion of the generated summaries. Additionally, participants mentioned that the summary can be a good starting point for new editors.

Acknowledgements

This research was supported by the EPSRC-funded project Data Stories under grant agreement number EP/P025676/1; and DSTL under grant number DSTLX-1000094186. We would like to express our gratitude to the Wikipedia communities involved in this study for embracing our research and particularly to the participants of the interviews. Hady Elsahar has been an essential part of this work, discussing the ideas with us from the beginning. We thank him for his support and collaboration on this and previous works on the topic of ArticlePlaceholder.

Appendix A. Guideline for the semi-structured interview

– Opening/Introduction

- * I am Lucie, a researcher at the University of Southampton and I work on this project as part of my PhD research, collaborating with Pavlos Vougiouklis and Elena Simperl.
- * Before I start about the content, I want to ask for your consent to participate in this study, according to the Ethics Committee of the University of Southampton. We will treat your data confidentiality and it will only be stored on the password-protected computer of the researchers. You have the option to withdraw, but we will have to ask you to do that up to 2 weeks after today.
- * We will use the results anonymized to provide insights into the editing of Wikipedia editors and publish the results of the study to a research venue. This experiment will observe your interaction with text and how you edit Wikipedia.
- * Do you agree to participate in this study?

– Demographic Questions

- * Do you read Wikipedia?
 - * In which language do you usually read Wikipedia?
 - * Do you search topics on Wikipedia or search engines (google)?
 - * If you can't find a topic on Wikipedia, what do you do?
- * Do you edit Wikipedia?
 - * What is your Wikimedia/Wikipedia username?
 - * Which Wikipedia do you mainly contribute to?
 - * How long have you contributed to Wikipedia?
 - * When you edit, what topics do you choose?
 - Topics you are interested in or topics that you think that is needed?
 - How do you decide, what is interesting/needed?
- * How do you usually start editing?
 - Where do you look up information? (for this topic specifically, in general)
 - Do you draft points and then write text or write the text first?

- * Have you heard of Wikidata?
 - * What is Wikidata?
 - * What is the relationship between Wikidata and Wikipedia?
 - * Have you edited it before?
 - * How long are you contributing to Wikidata?
- Description of AP
 - * This project is based on the ArticlePlaceholder.
 - * The idea is if there is no information about a topic, a user can search on Wikipedia and still get the information on the topic, that is available on Wikidata.
 - * We do not create stub articles, everything is displayed dynamically.
 - * Have you heard about the ArticlePlaceholder?
- Reading the layout of the AP
 - * Is the topic familiar to you? If so, how?
 - * If you look at the page, what information do you look at first? (If you want you can point it out with your mouse)
 - * What information do you look at after that? (In which order?)
 - * What part do you think is taken directly from Wikidata and what is from other sources? Which sources?
 - * What information is particularly helpful on this page?
 - * What do you think of the text in addition to the rest of the information?
- After editing sentence
 - * Where do you start in this case?
 - * What information do you miss when editing?
 - * Would you prefer to have more or less information given? What type of information?
 - * Would you prefer a longer text, even if it has a similar amount of missing information as this one?
- Closing Questions
 - * What we do in this project is to generate an introductory sentence from Wikidata triples. We train on this language Wikipedia.
 - * What impact do you think this project can have for you as a reader?
 - * Do you believe this project will have an impact on your editing?
 - * What does it change?
 - * Any questions from the interviewee?

Appendix B. Sentences created by the editors in the interviews and English translations

Language	Sentence by editor	Translation to English
Swedish	Marrakech (arabiska: مراکش, tamazight: ⵎⴰⵔⴰⴽⵉⵝ) är en stad i Marocko med 928 850 invånare (2014). Staden grundades 1062. Staden ligger 468 meter över havet.	Marrakech (Arabic: مراکش, tamazight: ⵎⴰⵔⴰⴽⵉⵝ) is a city in Morocco with 928,850 inhabitants (2014). The city was founded in 1062. The city is located 468 meters above sea level.
Swedish	Marrakech (arabiska: مراکش) är en stad i sydvästra Marocko, vid foten av Atlasbergen Marrakech tillhör sedan 2 mars 1956 Marocko and tillhörde innan dess Frankrike (30 mars 1912 - 2 mars 1956).	Marrakech (Arabic: مراکش) is a city in southwestern Morocco, at the foot of the Atlas Mountains Marrakech has belonged to Morocco since March 2, 1956 and before that belonged to France (March 30, 1912 - March 2, 1956).
Arabic	مُرَّاكُش (بالأمازيغية: <rare>, التسمية المحلية بالأمازيغية وسط الأطلس:) هي مدينة مغربية تقع شمال المغرب. تم إنشاء هذه المدينة في عام 1062. ترتفع المدينة عن مستوى سطح البحر بـ468 متر. عدد سكان المدينة 928-850 نسمة حسب الإحصائيات في 1 يناير 2014. تبلغ مساحة مراکش 230 كيلومتر مربع.	Marrakesh (in Berber: <rare>, the local name in Berber, in the middle of the Atlas) is a Moroccan city located in the north of Morocco. This city was established in the year 1062. The city is 468 meters above sea level. The population of the city is 928,850, according to statistics on January 1, 2014. Marrakech has an area of 230 square kilometres.
Arabic	مُرَّاكُش (بالأمازيغية: <rare>) هي مدينة مغربية تقع شمال البلاد. مراکش معروفة كمدينة سياحية، حيث يذهب العديد من العرب والمغاربة والأوروبيين للمدينة، تعتبر المدينة من أكبر المدن المغربية.	Marrakesh (Berber: <rare>) is a Moroccan city located in the north of the country. Marrakech is known as a tourist city, where many Arabs, Moroccans and Europeans go to the city. The city is considered one of the largest in Morocco.
Arabic	مُرَّاكُش (بالأمازيغية: ⵎⴰⵔⴰⴽⵉⵝ) هي مدينة مغربية تقع شمال البلاد. هي مدينة مغربية تقع شمال البلاد. تأسست في 1062. عدد السكان 968.850. مساحتها 230.	Marrakech (Amazigh: ⵎⴰⵔⴰⴽⵉⵝ) is a Moroccan city located in the north of the country. It is a Moroccan city located in the north of the country. Founded in 1062. Population 968,850. Its chastity is 230.
Arabic	مُرَّاكُش (بالأمازيغية: <rare>) هي مدينة مغربية تقع شمال البلاد.	Marrakesh (in Berber: <a missing word>) is a Moroccan city located in the north of the country.
Persian	شهر مراکش (به بربری: <rare>) یکی از شهرهای کشور مراکش و مرکز استان مراکش <rare> است که در سال ۱۰۶۲ بنیان‌گذاری شده‌است و حدود یک میلیون نفر جمعیت دارد. این شهر قبلاً بخشی از فرانسه تا سال ۱۹۵۶ بوده و بعد از استقلال مراکش، در این کشور قرار دارد.	The city of Morocco (Berber: <rare>) is one of the cities of Morocco and the capital of the province of Morocco <rare>, which was founded in 1062 and has a population of about one million people. The city was previously part of France until 1956 and is located in Morocco after independence.
Ukrainian	Марракеш (араб. مراکش) — важливе імперське місто в Марокко, розташованого біля підніжжя гір.	Marrakech (Arabic: مراکش) is an important imperial city in Morocco, located at the foot of the mountains.
Indonesian	Marrakesh (Arab: <rare>) adalah sebuah kota yang terletak di bagian barat daya negara Maroko <rare>.	Marrakesh (Arabic: <rare>) is a city located in the southwestern part of Morocco <rare>.
Hebrew	מרקש (בערבית: مراکش; בתאזיגית: ⵎⴰⵔⴰⴽⵉⵝ) היא עיר מדברית בדרום מערב מרוקו למרגלות הרי האטלס.	Marrakesh (Arabic: مراکش; Tamazigat: ⵎⴰⵔⴰⴽⵉⵝ) is a desert city in southwestern Morocco at the foot of the Atlas Mountains.

References

- [1] B.T. Adler, K. Chatterjee, L. de Alfaro, M. Faella, I. Pye and V. Raman, Assigning trust to Wikipedia content, in: *Proceedings of the 2008 International Symposium on Wikis, 2008*, Porto, Portugal, September 8–10, 2008, 2008. doi:[10.1145/1822258.1822293](https://doi.org/10.1145/1822258.1822293).
- [2] B.T. Adler and L. de Alfaro, A content-driven reputation system for the Wikipedia, in: *Proceedings of the 16th International Conference on World Wide Web, WWW 2007*, Banff, Alberta, Canada, May 8–12, 2007, 2007, pp. 261–270. doi:[10.1145/1242572.1242608](https://doi.org/10.1145/1242572.1242608).
- [3] I. Alegria, U. Cabezón, U.F. de Betoño, G. Labaka, A. Mayor, K. Sarasola and A. Zubiaga, Reciprocal enrichment between Basque Wikipedia and machine translation, in: *The People's Web Meets NLP, Collaboratively Constructed Language Resources*, 2013, pp. 101–118. doi:[10.1007/978-3-642-35085-6_4](https://doi.org/10.1007/978-3-642-35085-6_4).
- [4] G. Angeli, P. Liang and D. Klein, A simple domain-independent probabilistic approach to generation, in: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 502–512, <http://dl.acm.org/citation.cfm?id=1870658.1870707>.
- [5] J. Antin and C. Cheshire, Readers are not free-riders: Reading as a form of participation on Wikipedia, in: *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, CSCW 2010*, Savannah, Georgia, USA, February 6–10, 2010, 2010, pp. 127–130. doi:[10.1145/1718918.1718942](https://doi.org/10.1145/1718918.1718942).
- [6] P. Bao, B. Hecht, S. Carton, M. Quaderi, M. Horn and D. Gergle, Omnipedia: Bridging the Wikipedia language gap, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2012, pp. 1075–1084. doi:[10.1145/2207676.2208553](https://doi.org/10.1145/2207676.2208553).
- [7] J.A. Bateman, C. Matthiessen and L. Zeng, Multilingual natural language generation for multilingual software: A functional linguistic approach, *Appl. Artif. Intell.* **13**(6) (1999), 607–639. doi:[10.1080/088395199117289](https://doi.org/10.1080/088395199117289).
- [8] S. Bird, NLTK: The natural language toolkit, in: *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Proceedings of the Conference, Sydney, Australia, 17–21 July 2006, 2006.
- [9] J.E. Blumenstock, Size matters: Word count as a measure of quality on Wikipedia, in: *Proceedings of the 17th International Conference on World Wide Web, WWW 2008*, Beijing, China, April 21–25, 2008, 2008, pp. 1095–1096. doi:[10.1145/1367497.1367673](https://doi.org/10.1145/1367497.1367673).
- [10] N. Bouayad-Agha, G. Casamayor and L. Wanner, Natural language generation in the context of the semantic web, *Semantic Web* **5**(6) (2014), 493–513. doi:[10.3233/SW-130125](https://doi.org/10.3233/SW-130125).
- [11] A. Chisholm, W. Radford and B. Hachey, Learning to generate one-sentence biographies from Wikidata, in: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 1 (Long Papers), Association for Computational Linguistics, Valencia, Spain, 2017, pp. 633–642.
- [12] K. Cho, B. van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, October 25–29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, A. Moschitti, B. Pang and W. Daelemans, eds, ACL, 2014, pp. 1724–1734. doi:[10.3115/v1/d14-1179](https://doi.org/10.3115/v1/d14-1179).
- [13] P.D. Clough, R.J. Gaizauskas, S.S.L. Piao and Y. Wilks, METER: MEasuring TEXT reuse, in: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, July 6–12, 2002, Philadelphia, PA, USA, 2002, pp. 152–159.
- [14] B. Collier and J. Bear, Conflict, criticism, or confidence: An empirical examination of the gender gap in Wikipedia contributions, in: *CSCW '12 Computer Supported Cooperative Work*, Seattle, WA, USA, February 11–15, 2012, 2012, pp. 383–392. doi:[10.1145/2145204.2145265](https://doi.org/10.1145/2145204.2145265).
- [15] L. Dong and M. Lapata, Language to logical form with neural attention, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, August 7–12, 2016, Berlin, Germany, Vol. 1 (Long Papers), 2016, <http://aclweb.org/anthology/P/P16/P16-1004.pdf>.
- [16] X. Du, J. Shao and C. Cardie, Learning to ask: Neural question generation for reading comprehension, in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, Vancouver, Canada, July 30–August 4, Vol. 1 (Long Papers), 2017, pp. 1342–1352. doi:[10.18653/v1/P17-1123](https://doi.org/10.18653/v1/P17-1123).
- [17] D. Duma and E. Klein, Generating natural language from linked data: Unsupervised template extraction, in: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, Association for Computational Linguistics, Potsdam, Germany, 2013, pp. 83–94, <http://www.aclweb.org/anthology/W13-0108>.
- [18] B. Ell and A. Harth, A language-independent method for the extraction of RDF verbalization templates, in: *INLG 2014 – Proceedings of the Eighth International Natural Language Generation Conference, Including Proceedings of the INLG and SIGDIAL 2014 Joint Session*, 19–21 June 2014, Philadelphia, PA, USA, 2014, pp. 26–34.
- [19] H. ElSahar, C. Gravier and F. Laforest, Zero-shot question generation from knowledge graphs for unseen predicates and entity types, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*, New Orleans, Louisiana, USA, June 1–6, 2018, Vol. 1 (Long Papers), 2018, pp. 218–228, <https://aclanthology.info/papers/N18-1020/n18-1020>.
- [20] D. Galanis and I. Androutsopoulos, Generating multilingual descriptions from linguistically annotated OWL ontologies: The NaturalOWL system, in: *Proceedings of the Eleventh European Workshop on Natural Language Generation*, Association for Computational Linguistics, 2007, pp. 143–146.
- [21] C. Gardent, A. Shimorina, S. Narayan and L. Perez-Beltrachini, The WebNLG challenge: Generating text from RDF data, in: *Proceedings of the 10th International Conference on Natural Language Generation*, Association for Computational Linguistics, 2017, pp. 124–133, <http://aclweb.org/anthology/W17-3518>.
- [22] S. Gehrmann, F. Dai, H. Elder and A. Rush, End-to-end content and plan selection for data-to-text generation, in: *Proceedings of the 11th International Conference on Natural Language Generation*, Association for Computational Linguistics, Tilburg University, The Netherlands, 2018, pp. 46–56, <https://www.aclweb.org/anthology/W18-6505>.

- [23] R.S. Geiger and D. Ribes, The work of sustaining order in Wikipedia: The banning of a vandal, in: *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, CSCW 2010*, Savannah, Georgia, USA, February 6–10, 2010, 2010, pp. 117–126. doi:[10.1145/1718918.1718941](https://doi.org/10.1145/1718918.1718941).
- [24] M. Graham, B. Hogan, R.K. Straumann and A. Medhat, Uneven geographies of user-generated information: Patterns of increasing informational poverty, *Annals of the Association of American Geographers* **104**(4) (2014), 746–764. doi:[10.1080/00045608.2014.910087](https://doi.org/10.1080/00045608.2014.910087).
- [25] S. Green, J. Heer and C.D. Manning, The efficacy of human post-editing for language translation, in: *2013 ACM SIGCHI Conference on Human Factors in Computing Systems, CHI '13*, Paris, France, April 27–May 2, 2013, 2013, pp. 439–448. doi:[10.1145/2470654.2470718](https://doi.org/10.1145/2470654.2470718).
- [26] S.A. Hale, Multilinguals and Wikipedia editing, in: *ACM Web Science Conference, WebSci '14*, Bloomington, IN, USA, June 23–26, 2014, 2014, pp. 99–108. doi:[10.1145/2615569.2615684](https://doi.org/10.1145/2615569.2615684).
- [27] S.A. Hale, Cross-language Wikipedia editing of Okinawa, Japan, in: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015*, Seoul, Republic of Korea, April 18–23, 2015, 2015, pp. 183–192. doi:[10.1145/2702123.2702346](https://doi.org/10.1145/2702123.2702346).
- [28] N. Halko, P. Martinsson and J.A. Tropp, Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, *SIAM Review* **53**(2) (2011), 217–288. doi:[10.1137/090771806](https://doi.org/10.1137/090771806).
- [29] A. Hartley and C. Paris, Multilingual document production from support for translating to support for authoring, *Machine Translation* **12**(1) (1997), 109–129. doi:[10.1023/A:1007986908015](https://doi.org/10.1023/A:1007986908015).
- [30] K. Heafield, I. Pouzyrevsky, J.H. Clark and P. Koehn, Scalable modified Kneser–Ney language model estimation, in: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013*, 4–9 August 2013, Sofia, Bulgaria, Vol. 2 (Short Papers), 2013, pp. 690–696.
- [31] B. Hecht and D. Gergle, The tower of Babel meets Web 2.0: User-generated content and its applications in a multilingual context, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2010, pp. 291–300. doi:[10.1145/1753326.1753370](https://doi.org/10.1145/1753326.1753370).
- [32] B.J. Hecht, The mining and application of diverse cultural perspectives in user-generated content, PhD thesis, Northwestern University, 2013.
- [33] C. Hu, B.B. Bederson, P. Resnik and Y. Kronrod, MonoTrans2: A new human computation system to support monolingual translation, in: *Proceedings of the International Conference on Human Factors in Computing Systems, CHI 2011*, Vancouver, BC, Canada, May 7–12, 2011, 2011, pp. 1133–1136. doi:[10.1145/1978942.1979111](https://doi.org/10.1145/1978942.1979111).
- [34] P. Isola, J. Zhu, T. Zhou and A.A. Efros, Image-to-image translation with conditional adversarial networks, in: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Honolulu, HI, USA, July 21–26, 2017, 2017, pp. 5967–5976. doi:[10.1109/CVPR.2017.632](https://doi.org/10.1109/CVPR.2017.632).
- [35] T. Joachims, A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization, in: *Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997)*, Nashville, Tennessee, USA, July 8–12, 1997, 1997, pp. 143–151.
- [36] L. Kaffee, H. ElSahar, P. Vougiouklis, C. Gravier, F. Laforest, J.S. Hare and E. Simperl, 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, NAACL-HLT*, New Orleans, Louisiana, USA, June 1–6, 2018, Vol. 2 (Short Papers), 2018, pp. 640–645. <https://aclanthology.info/papers/N18-2101/n18-2101>.
- [37] L. Kaffee, H. ElSahar, P. Vougiouklis, C. Gravier, F. Laforest, J.S. Hare and E. Simperl, Mind the (language) gap: Generation of multilingual Wikipedia summaries from Wikidata for ArticlePlaceholders, in: *The Semantic Web – 15th International Conference, ESWC 2018*, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings, Lecture Notes in Computer Science, Vol. 10843, Springer, 2018, pp. 319–334. doi:[10.1007/978-3-319-93417-4_21](https://doi.org/10.1007/978-3-319-93417-4_21).
- [38] L. Kaffee, K.M. Endris and E. Simperl, When humans and machines collaborate: Cross-lingual label editing in Wikidata, in: *Proceedings of the 15th International Symposium on Open Collaboration, OpenSym 2019*, Skövde, Sweden, August 20–22, 2019, B. Lundell, J. Gamalielsson, L. Morgan and G. Robles, eds, ACM, 2019, pp. 16:1–16:9. doi:[10.1145/3306446.3340826](https://doi.org/10.1145/3306446.3340826).
- [39] L.-A. Kaffee, A. Piscopo, P. Vougiouklis, E. Simperl, L. Carr and L. Pintscher, A glimpse into Babel: An analysis of multilinguality in Wikidata, in: *Proceedings of the 13th International Symposium on Open Collaboration*, ACM, 2017, p. 14.
- [40] L.-A. Kaffee and E. Simperl, Analysis of editors’ languages in Wikidata, in: *Proceedings of the 14th International Symposium on Open Collaboration, OpenSym 2018*, Paris, France, August 22–24, 2018, 2018, pp. 21:1–21:5. doi:[10.1145/3233391.3233965](https://doi.org/10.1145/3233391.3233965).
- [41] L.A. Kaffee, Generating ArticlePlaceholders from Wikidata for Wikipedia: Increasing access to free and open knowledge, 2016.
- [42] R.I. Kittredge, A. Polguère and E. Goldberg, Synthesizing weather forecasts from formatted data, in: *Proceedings of the 11th International Conference on Computational Linguistics, COLING '86*, Bonn, Germany, August 25–29, 1986, Institut für angewandte Kommunikations- und Sprachforschung e.V. (IKS), Poppelsdorfer Allee 47, Bonn, Germany, 1986, pp. 563–565. <https://www.aclweb.org/anthology/C86-1132/>.
- [43] A. Kittur, B. Suh and E.H. Chi, Can you ever trust a Wiki?: Impacting perceived trustworthiness in Wikipedia, in: *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work, CSCW 2008*, San Diego, CA, USA, November 8–12, 2008, 2008, pp. 477–480. doi:[10.1145/1460563.1460639](https://doi.org/10.1145/1460563.1460639).
- [44] R. Koncel-Kedziorski, D. Bekal, Y. Luan, M. Lapata and H. Hajishirzi, Text generation from knowledge graphs with graph transformers, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 2284–2293. <https://www.aclweb.org/anthology/N19-1238>. doi:[10.18653/v1/N19-1238](https://doi.org/10.18653/v1/N19-1238).
- [45] I. Konstas and M. Lapata, A global model for concept-to-text generation, *J. Artif. Int. Res.* **48**(1) (2013), 305–346. <http://dl.acm.org/citation.cfm?id=2591248.2591256>.
- [46] C. Kramsch and H. Widdowson, *Language and Culture*, Oxford University Press, 1998.

- [47] G.M. Kruijff, E. Teich, J.A. Bateman, I. Kruijff-Korbayová, H. Skoumalová, S. Sharoff, E.G. Sokolova, T. Hartley, K. Staykova and J. Hana, Multilinguality in a text generation system for three Slavic languages, in: *COLING 2000, 18th International Conference on Computational Linguistics, Proceedings of the Conference, 2 Volumes*, July 31–August 4, Universität des Saarlandes, Saarbrücken, Germany, Morgan Kaufmann, 2000, pp. 474–480, <https://www.aclweb.org/anthology/C00-1069/>.
- [48] S. Kuznetsov, Motivations of contributors to Wikipedia, *SIGCAS Computers and Society* **36**(2) (2006), 1. doi:10.1145/1215942.1215943.
- [49] E. Lagoudaki, Translation editing environments, in: *MT Summit XII: Workshop on Beyond Translation Memories*, 2009.
- [50] A. Lavie and A. Agarwal, METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments, in: *Proceedings of the Second Workshop on Statistical Machine Translation, StatMT '07*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2007, pp. 228–231. doi:10.3115/1626355.1626389.
- [51] R. Lebrecht, D. Grangier and M. Auli, 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, EMNLP 2016*, Austin, Texas, USA, November 1–4, 2016, 2016, pp. 1203–1213.
- [52] F. Lemmerich, D. Sáez-Trumper, R. West and L. Zia, Why the world reads Wikipedia: Beyond English speakers, in: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019*, Melbourne, VIC, Australia, February 11–15, 2019, 2019, pp. 618–626. doi:10.1145/3289600.3291021.
- [53] W.D. Lewis and P. Yang, Building MT for a severely under-resourced language: White Hmong, Association for Machine Translation in the Americas, 2012.
- [54] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, Vol. 8, Barcelona, Spain, 2004.
- [55] T. Liu, K. Wang, L. Sha, B. Chang and Z. Sui, Table-to-Text Generation by Structure-Aware Seq2seq Learning, 2018, <https://www.aaii.org/ocs/index.php/AAAI/AAAI18/paper/view/16599>.
- [56] T. Lucassen and J.M. Schraagen, Trust in Wikipedia: How users trust information from an unknown source, in: *Proceedings of the 4th ACM Workshop on Information Credibility on the Web, WICOW 2010*, Raleigh, North Carolina, USA, April 27, 2010, 2010, pp. 19–26. doi:10.1145/1772938.1772944.
- [57] T. Luong, H. Pham and C.D. Manning, Effective approaches to attention-based neural machine translation, in: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 1412–1421, <http://aclweb.org/anthology/D15-1166>. doi:10.18653/v1/D15-1166.
- [58] T. Luong, I. Sutskever, Q.V. Le, O. Vinyals and W. Zaremba, Addressing the rare word problem in neural machine translation, in: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015*, July 26–31, 2015, Beijing, China, Vol. 1 (Long Papers), 2015, pp. 26–31.
- [59] H. Mei, M. Bansal and M.R. Walter, What to talk about and how? Selective generation using LSTMs with coarse-to-fine alignment, in: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, San Diego, California, 2016, pp. 720–730, <http://www.aclweb.org/anthology/N16-1086>.
- [60] C. Mellish and R. Dale, Evaluation in the context of natural language generation, *Computer Speech & Language* **12**(4) (1998), 349–373. doi:10.1006/csla.1998.0106.
- [61] M. Mintz, S. Bills, R. Snow and D. Jurafsky, Distant supervision for relation extraction without labeled data, in: *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 2–7 August 2009, Singapore, 2009, pp. 1003–1011.
- [62] J.T. Morgan, S. Bouterse, H. Walls and S. Stierch, Tea and sympathy: Crafting positive new user experiences on Wikipedia, in: *Computer Supported Cooperative Work, CSCW 2013*, San Antonio, TX, USA, February 23–27, 2013 2013, pp. 839–848. doi:10.1145/2441776.2441871.
- [63] S. Narayan, J. Orlowitz, J.T. Morgan, B.M. Hill and A.D. Shaw, The Wikipedia adventure: Field evaluation of an interactive tutorial for new users, in: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017*, Portland, OR, USA, February 25–March 1, 2017, 2017, pp. 1785–1799, <http://dl.acm.org/citation.cfm?id=2998307>.
- [64] A.-C. Ngonga Ngomo, L. Bühmann, C. Unger, J. Lehmann and D. Gerber, Sorry, I don't speak SPARQL – Translating SPARQL queries into natural language, in: *Proceedings of the 22nd International Conference on World Wide Web*, ACM, 2013, pp. 977–988. doi:10.1145/2488388.2488473.
- [65] K.A. Panciera, A. Halfaker and L.G. Terveen, Wikipedians are born, not made: A study of power editors on Wikipedia, in: *Proceedings of the 2009 International ACM SIGGROUP Conference on Supporting Group Work, GROUP 2009*, Sanibel Island, Florida, USA, May 10–13, 2009, 2009, pp. 51–60. doi:10.1145/1531674.1531682.
- [66] K. Papineni, S. Roukos, T. Ward and W. Zhu, Bleu: A method for automatic evaluation of machine translation, in: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, July 6–12, 2002, Philadelphia, PA, USA, 2002, pp. 311–318.
- [67] M.J. Pat Wu, State of connectivity 2015: A report on global internet access, 2016, <http://newsroom.fb.com/news/2016/02/state-of-connectivity-2015-a-report-on-global-internet-access/>.
- [68] P. Pirolli, E. Wöllny and B. Suh, So you know you're getting the best possible information: A tool that increases Wikipedia credibility, in: *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009*, Boston, MA, USA, April 4–9, 2009, 2009, pp. 1505–1508. doi:10.1145/1518701.1518929.
- [69] Y. Pochampally, K. Karlapalem and N. Yarrabally, Semi-supervised automatic generation of Wikipedia articles for named entities, in: *Wiki@ ICWSM*, 2016.

- [70] M. Potthast, T. Gollub, M. Hagen, J. Kiesel, M. Michel, A. Oberländer, M. Tippmann, A. Barrón-Cedeño, P. Gupta, P. Rosso and B. Stein, Overview of the 4th international competition on plagiarism detection, in: *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes*, Rome, Italy, September 17–20, 2012, 2012.
- [71] E. Reiter, Natural language generation, in: *The Handbook of Computational Linguistics and Natural Language Processing*, Wiley-Blackwell, 2010, pp. 574–598, Chapter 20. ISBN 9781444324044. doi:10.1002/9781444324044.ch20.
- [72] E. Reiter and A. Belz, An investigation into the validity of some metrics for automatically evaluating natural language generation systems, *Comput. Linguist.* **35**(4) (2009), 529–558. doi:10.1162/coli.2009.35.4.35405.
- [73] E. Reiter, R. Robertson and S. Sripada, Acquiring correct knowledge for natural language generation, *J. Artif. Int. Res.* **18** (2003), 491–516, <https://jair.org/index.php/jair/article/view/10332>. doi:10.1613/jair.1176.
- [74] A. Rohrbach, L.A. Hendricks, K. Burns, T. Darrell and K. Saenko, Object hallucination in image captioning, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, October 31–November 4, 2018, 2018, pp. 4035–4045, <https://aclanthology.info/papers/D18-1437/d18-1437>. doi:10.18653/v1/D18-1437.
- [75] A.M. Rush, S. Chopra and J. Weston, A neural attention model for abstractive sentence summarization, in: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, Lisbon, Portugal, September 17–21, 2015, 2015, pp. 379–389.
- [76] C. Sauper and R. Barzilay, 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*, Vol. 1, Association for Computational Linguistics, 2009, pp. 208–216.
- [77] A. See, P.J. Liu and C.D. Manning, Get to the point: Summarization with pointer-generator networks, in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vol. 1 (Long Papers), Association for Computational Linguistics, 2017, pp. 1073–1083, <http://www.aclweb.org/anthology/P17-1099>. doi:10.18653/v1/P17-1099.
- [78] I.V. Serban, A. García-Durán, Ç. Gülçehre, S. Ahn, S. Chandar, A.C. Courville and Y. Bengio, Generating factoid questions with recurrent neural networks: The 30M factoid question–answer corpus, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, August 7–12, 2016, Berlin, Germany, Vol. 1 (Long Papers), 2016.
- [79] A. Sleimi and C. Gardent, Generating paraphrases from DBpedia using deep learning, in: *Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016)*, Association for Computational Linguistics, 2016, pp. 54–57, <http://www.aclweb.org/anthology/W16-3511>.
- [80] T. Steiner, Bots vs. Wikipedians, anons vs. logged-ins (redux): A global study of edit activity on Wikipedia and Wikidata, in: *Proceedings of the International Symposium on Open Collaboration, OpenSym '14*, ACM, New York, NY, USA, 2014, pp. 25:1–25:7. ISBN 978-1-4503-3016-9. doi:10.1145/2641580.2641613.
- [81] X. Sun and C. Mellish, An experiment on free generation from single RDF triples, in: *Proceedings of the Eleventh European Workshop on Natural Language Generation*, Association for Computational Linguistics, 2007, pp. 105–108.
- [82] I. Sutskever, O. Vinyals and Q.V. Le, Sequence to sequence learning with neural networks, in: *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence and K.Q. Weinberger, eds, Curran Associates, Inc., 2014, pp. 3104–3112.
- [83] J. Voss, Measuring Wikipedia, in: *Proceedings of 10th International Conference of the International Society for Scientometrics and Informetrics*, 2005.
- [84] P. Vougiouklis, H. Elshahar, L.-A. Kaffee, C. Gravier, F. Laforest, J. Hare and E. Simperl, Neural Wikipedian: Generating textual summaries from knowledge base triples, *Journal of Web Semantics* (2018), <http://www.sciencedirect.com/science/article/pii/S1570826818300313>. doi:10.1016/j.websem.2018.07.002.
- [85] P. Vougiouklis, E. Maddalena, J.S. Hare and E. Simperl, Point at the triple: Generation of text summaries from knowledge base triples, *J. Artif. Int. Res.* **69** (2020), 1–31. doi:10.1613/jair.1.11694.
- [86] D. Vrandečić and M. Krötzsch, Wikidata: A free collaborative knowledgebase, *Communications of the ACM* **57**(10) (2014), 78–85. doi:10.1145/2629489.
- [87] L. Wanner, B. Bohnet, N. Bouayad-Agha, F. Lareau and D. Nicklaß, Marquis: Generation of user-tailored multilingual air quality bulletins, *Applied Artificial Intelligence* **24**(10) (2010), 914–952. doi:10.1080/08839514.2010.529258.
- [88] S. Williams and E. Reiter, Generating basic skills reports for low-skilled readers, *Natural Language Engineering* **14**(4) (2008), 495–525. doi:10.1017/S1351324908004725.
- [89] M.J. Wise, YAP3: Improved detection of similarities in computer program and other texts, *ACM SIGCSE Bulletin* **28**(1) (1996), 130–134. doi:10.1145/236462.236525.
- [90] S. Wiseman, S. Shieber and A. Rush, Challenges in data-to-document generation, in: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, Copenhagen, Denmark, 2017, pp. 2253–2263, <https://www.aclweb.org/anthology/D17-1239>.
- [91] S. Yeh, H. Huang and H. Chen, Precise description generation for knowledge base entities with local pointer network, in: *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 2018, pp. 214–221. doi:10.1109/WI.2018.00-87.