Capturing meaning: Toward an abstract Wikipedia

Denny Vrandečić

Google
vrandecic@google.com

Abstract. Semantic Web languages allow to express ontologies and knowledge bases in a way meant to be particularly amenable to the Web. Ontologies formalize the shared understanding of a domain. But the most expressive and widespread languages that we know of are human natural languages, and the largest knowledge base we have is the wealth of text written in human languages. This paper looks for a path to bridge the gap between knowledge representation languages such as OWL and human natural languages such as English. We propose a project to simultaneously expose that gap, allow to collaborate on closing it, make progress widely visible, and is highly attractive and valuable in its own right: a Wikipedia written in an abstract language to be rendered into any natural language on request. This would make current Wikipedia editors about 100x more productive, and increase the content of Wikipedia by 10x. For billions of users this will unlock knowledge they currently do not have access to.

Keywords: Semantics · Multilingual · Abstract language.

1 Wikipedia today

Wikipedia is one of the most important sources of knowledge today. Following its vision to allow everyone to share in the sum of all knowledge, it aims to create comprehensive and current encyclopedias anyone can read and contribute to.

In order to read Wikipedia it has to be available in the reader's language. Wikipedia is available in 300 languages. But content is unevenly distributed:

First, English has 5.6 million articles, 60 languages have over 100,000 articles – but many languages are tiny: Zulu has 901, Samoan 797, and Cree 131 articles.

Second, topic coverage has a surprisingly low overlap. English and German are the two most active Wikipedias. English has 5.6 million, German 2.1 million articles – but only 1.1 million of the German topics are also available in English. Only 100,000 topics are common between the top ten most active Wikipedias.

Third, the comprehensiveness of individual articles differ between languages: whereas the English Wikipedia has a single line on the Port of Călăraşi, the Romanian Wikipedia offers several paragraphs, including a picture. "Local" Wikipedias often have information that is missing from others.

There are currently Wikipedia articles on 17.9 million topics. Having them available in all 300 languages of Wikipedia would lead to more than 5.3 billion articles. This dwarfs the number of 50 million articles that actually exist – less than one percent of that goal. And this does not take into consideration the effort required to maintain these articles: there are about 69,000 active contributors. 31,000 are active on the English Wikipedia, followed by German with 5,500. Only eleven Wikipedias have thousand or more active editors. More than half of Wikipedias have less than ten active editors. Completing and maintaining an encyclopedia with only ten volunteers is ambitious.

The underlying issue is that the current size of the problem is the number of topics *multiplied* with the number of languages. In the following we suggest a solution that reduces it so that the size of the problem is essentially the number of topics *added* to the number of languages.

2 Recognizing the formal language gap

Take the first two paragraphs of a Wikipedia article, extract the content into an RDF graph, randomize the order of the triples, generate a natural language text from the graph, and compare the result with the original text.[1]

This round-tripping will expose the gap between natural language and RDF. Let us take a sentence from the Wikipedia article about the Rosetta Stone:[2]

"It was the first Ancient Egyptian bilingual text recovered in modern times, and it aroused widespread public interest with its potential to decipher this previously untranslated hieroglyphic language."

Just to note one issue (among many): the sentence states that Ancient Egyptian is hieroglyphic. This is not knowledge about the Rosetta Stone, but about Ancient Egyptian, and it has been included in the article on the Rosetta Stone deliberately to emphasize the challenge of translating the language.

We think that a naïve RDF representation of the declarative knowledge expressed in the text is not the best choice to meet the challenge posed here, but that we need an alternative representation that is capable of capturing the narrative, redundancy, and organization of the text more closely.

3 An abstract Wikipedia

The question which data to assemble in which order from a large declarative knowledge base is a hard one. We avoid it by leaving this task to the contributors. We sketch a solution consisting of three main components (see Fig. 1):

Content. The largest component is the Content: each individual article is represented by knowledge that captures the content of an article in an abstract, language-independent way. This captures paragraphs and sentences with their content, redundancies in the text, relevant knowledge about related topics, etc.

[1] The people extracting the RDF and generating the text should ideally be different.
[2] https://en.wikipedia.org/wiki/Rosetta_Stone

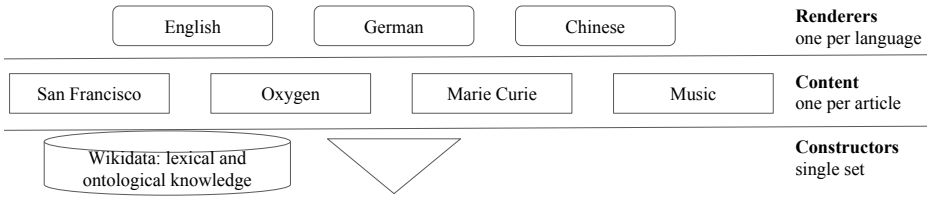

Fig. 1. The three components of the abstract Wikipedia are Renderers, Content, and Constructors. Additionally, Wikidata is used as background knowledge.

Constructors. The smallest component: this defines the 'language' used for the Content. If the Content is a series of function calls, these are the function definitions. If the Content are frame instantiations, these define the frames and slots. If the Content is akin to the ABox, then the Constructors are the TBox.

Renderers. The Renderers translate Content to natural language. They define for each Constructor how to represent it in natural language, using Wikidata as ontological and lexicographical background knowledge. The hope is that a Renderer requires only few contributors – maybe ten or less – in order to achieve interesting coverage. Then even a small group would be capable to create a comprehensive and up-to-date encyclopedia in their language.

This glosses over many issues such as agreement, saliency, how Renderers interact over Constructors in order to achieve high readability and proper anaphora creation, or different languages requiring different knowledge in the Content.[3]

4 Desiderata

The main feature of Wikipedia is that anyone can contribute to it. This must remain true of the abstract Wikipedia and its components.

Content must be easy to create, refine, and change. It will constitute the largest part, followed by lexical knowledge and, far behind, the Renderers and Constructors. If trade-offs between the components are needed, it should be taken into account how many contributors each component requires.

The set of Constructors has to be under control of and be extensible by the community. The Constructors and their individual slots, whether these slots are required or optional, etc., have to be editable and maintainable. The system must be able to deal with the Constructors evolving.

The Renderers can only scale to hundreds of languages if they are created by the community. But not everyone will need to be able to write Renderers: it could be a task which can only be done by contributors who dedicate the necessary time. Such a separation of concern exists today already: contributors

[3] An example: in English the sentence *"She is an actress."* requires to specify gender, whereas in Turkish it is unnecessary as there is no grammatical gender: *"O bir oyuncu."* Knowledge in the Content may be marked as core or supplemental.

have diverse skillsets and time commitments, from template developers and bot operators to casual contributors using the WYSIWIG interface.

Lexical knowledge must be easy to contribute. The Renderers will require large amounts of lexical knowledge. Fortunately, Wikidata has recently launched support for capturing and maintaining lexical knowledge.[4]

The system must support graceful degradation. Each language will grow at its own speed: some languages will have complete Renderers, others will be stale and incomplete. It is important that a missing lexicalization does not block a whole article. A sentence that renders in English as *"In 2013, the secretary of state was the first foreign official to visit the country since the revolution."* could degrade to *"The minister visited the country in 2013."* Parts of the Content could be marked as optional or to require other parts of the Content to be rendered, and thus allow to be dropped in case the language resources prove insufficient.

5 Unique advantages

There are several unique advantages to render the project more feasible:

- we only aim at a single genre of text, encyclopedias – not poems, fiction, etc.
- the exact text is less important as long as it contains the necessary content.
- where we rely on the exact text, we can have ways to quote verbatim.
- we can start with simple sentences and iterate. We won't achieve the expressivity of natural language, and yet may achieve very valuable goals.
- we do not need to understand and parse natural language, merely generate it, which is widely considered a much simpler task.
- the baseline is low. The first priority is about languages that currently have a small number of articles, many of which are out of date and incomplete.

6 Challenge

The abstract Wikipedia is a two-pronged approach: we want to push for new goals in knowledge representation, and we want to create an extremely valuable resource. We present a clearly defined, visible, and attractive, yet achievable challenge. It allows to collaborate and iterate on a system that captures more knowledge than current KR systems, and watch the system evolve and grow.

This paper offers no concrete solution. It introduces the challenge of an abstract Wikipedia with three questions: What knowledge representation is suitable? How are individual articles represented? How do the renderers work?

This is a call for developers and the research communities to answer the challenge of the abstract Wikipedia and demonstrate how the current state of the art in natural language generation, knowledge representation, and collaborative systems can work together to create a novel system that will enable everyone to share in the sum of all knowledge.

[4] see https://www.wikidata.org/wiki/Wikidata:Lexicographical_data