THE ROBOT WIKIPEDIAN

Background

Wikipedia is one of the most visited sites on the Web. Often an inquiry in any subject starts with reading a Wikipedia article followed by a further investigation. Wikipedia also supports more than 300 languages, making it one of the most accessible sources of information across languages.

Yet. Wikipedia is very unbalanced with respect to language coverage. For example, there are more than 6 million English articles, but the second largest language - German - has only 2.7 million. The majority of the other 300 languages has less than 100 000 articles per language. Wikipedia is also very culturally biassed. For example local customs, places and people are best described in the language of the community. Even when an article is available in a number of languages, it often contains different sometimes conflicting information in the different languages.

Abstract Wikipedia is a project from Wikimedia which aims to improve the situation by generating articles automatically from Wikidata, a database of computer readable facts, collected from Wikipedia.

Project description

Grammatical Framework (GF) is a programming language used by a community of researchers to describe the syntax, the morphology and the lexicons for a number of native languages. There are already resources for the majority of the European as well as many of the Asian and African languages.

The aim of the project is to evaluate the use of these language resources for the simultaneous generation of articles in several languages. To narrow down the scope of the project we will only focus on articles about famous people. For example a typical article may

contain facts like the name of the person, his nationality, profession, achievements, etc.

The goal is that the students should not actively program in GF but they will have to learn a bit of the API used to access the language resources. The API is available in Haskell, Python and Java. Web API for JavaScript is also available. The students should also learn how to access the data stored in Wikidata by using either SPARQL or the Entity Web API. The project should also provide a web interface where the newly generated articles can be explored.

The final product must be able to produce articles in all languages supported in GF, and the correctness will be checked at least for English and Swedish. Other languages may be considered if there are speakers of these languages in the group. For any other language the quality of the outcome will depend on the quality of the language resources in the framework.

Suggested reading material

The backend is based on Grammatical Framework (GF). Although an in-depth knowledge of GF is not necessary, a general background will be useful. There is a tutorial here:

http://www.grammaticalframework.org/doc/tutorial/gf-tutorial.html

By using the framework the community has built a library of resource grammars for different languages:

https://www.grammaticalframework.org/lib/doc/synopsis/index.html

In addition we will use the GF WordNet which is a large multilingual lexicon. Quick overview can be found here:

https://cloud.grammaticalframework.org/wordnet/gf-wordnet-help.html

Wikidata is an open-access graph database build by the Wikipedia community:

https://www.wikidata.org

The data can be queried by using SPARQL:

https://query.wikidata.org/

You can also fetch the data for a specific entity by accessing its unique URI. For example the data for Douglas Adams is available here:

https://www.wikidata.org/wiki/Special:EntityData/Q42.json (in JSON)
https://www.wikidata.org/wiki/Special:EntityData/Q42 (in HTML).

Target group

D, DV and IT
The project is suitable for students who are interested in natural language and web technologies.

Special prerequisites

Experience in JavaScript and Web technologies.

Proposal author

Krasimir Angelov