DATX02-22-27

Multilingual Text Robots for Abstract Wikipedia

Bachelor Thesis proposal, CSE, Chalmers and GU, 2021

Aarne Ranta, http://www.cse.chalmers.se/~aarne/ - author and supervisor

Background

Wikipedia is available in over 300 languages. However, most languages have only a few articles, and the corresponding articles in different languages can vary a lot in length and quality. The reason for this is obviously that human translation is a scarce resource. At the same time, automatic translation such as Google translate is not reliable enough for the task.

A possible way to solve the problem is **Natural Language Generation (NLG)**: algorithms that produce articles in different languages from a common data source automatically. An example of NLG are so-called **text robots**, which in fact have produced millions of Wikipedia articles, also in Swedish. Today's text robots are mostly limited to schematic texts in limited domains and single languages (e.g. Swedish lakes in Swedish). But a recent initiative from the Wikimedia foundation, **Abstract Wikipedia**, aims to take this to the next level, by developing multilingual NLG that can target, as its ultimate goal, over 300 languages simultaneously and enable rich and fluent language on multiple domains.

Project description

The proposed project is a part of the Abstract Wikipedia programme. Its aim is to investigate the use of **Grammatical Framework (GF)** as a technology for generating Wikipedia articles from **Wikidata**, which is a graph database containing the facts that the articles describe.

The final task of the project group is to generate a comprehensive set of Wikipedia articles in at least three languages. To do this, the group will select a domain of application, collect and organize relevant data, build a dictionary of appropriate terms, and write some software based on GF. Another goal is to evaluate the feasibility of the method: how easy will it be for a Wikipedia authors with a comparable education (BSc students) but no prior knowledge of GF to produce acceptable articles? The underlying GF programming will thus mostly come from the GF community and the supervisor, and be provided to the group via high-level APIs.

Suggested reading (YouTube films)

Abstract Wikipedia talk by Denny Vrandečić
Google tech talk about GF by Aarne Ranta
NLG for Abstract Wikipedia by Aarne Ranta

Prerequisites

Programming in Python is desirable. Functional programming, databases, and some web programming are a plus. It is also a plus if the participants together know many languages, but Swedish and English are the minimum.

Target groups

DV, D and IT