Automatically extracting the semantic network out of public services to support cities becoming smart cities

Joachim Van den Bogaert, Laurens Meeus, Alina Kramchaninova, Arne Defauw, Sara Szoc, Frederic Everaert, Koen Van Winckel, Anna Bardadym, Tom Vanallemeersch,
CrossLang NV, Kerkstraat 106, 9050 Gent, Belgium
{firstname.lastname}@crosslang.com

Abstract

The CEFAT4Cities project aims at creating a multilingual semantic interoperability layer for smart cities that allows users from all EU member states to interact with public services in their own language. The CEFAT4Cities processing pipeline transforms natural-language administrative procedures into machine-readable data using various multilingual natural-language processing techniques, such as semantic networks and machine translation, thus allowing for the development of more sophisticated and more user-friendly public services applications.

1 Introduction

To ease interaction with a city's administrative services, the creation of a chatbot is an easy and popular option, with many open-source platforms currently available. However, the main challenge lies in filling the bot with the right content, i.e. an accurate map that can predict exactly what a user is looking for, together with the relevant next steps to take or suggest. Normally, it takes a dedicated team of experienced editors to create such a "mind map" by collecting content and extracting all relevant information. This is a time-consuming process, even for a very limited use case. Imagine this for multiple use cases, in all EU languages administering a metropolitan area with citizens originating from all over the world.

The CEFAT4Cities project[1] aims at supporting cities in creating "semantic networks" of their public services, by building a processing pipeline that ingests legacy data from public services (from e.g. websites, administrative forms, existing applications) in multiple EU languages, and transforms this data into a network of connected services that can be used across applications and languages. By connecting the pipeline to the FIWARE Context Broker[2], the mind map is made available to any app or sensor within the smart-city IoT network.

2 Methodology

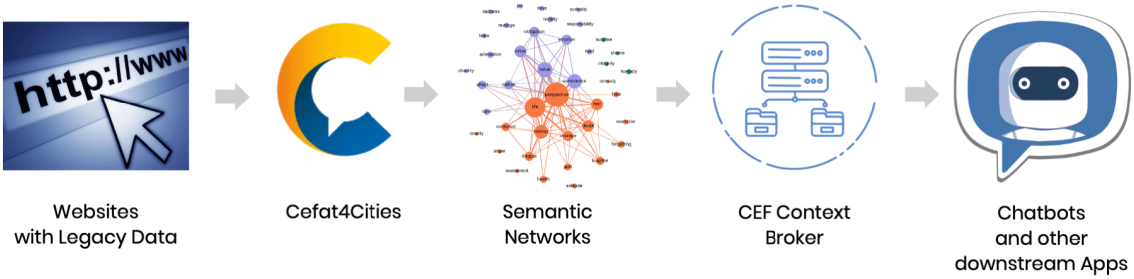
To create the semantic network of public services, CEFAT4Cities partners start from a few abstract templates that describe what a public service looks like (who can submit a form to get access to which service, providing which type of proof?) and what the interacting entities look like (are we dealing with an organisation or a citizen?). These abstract templates consist of nodes and links (hence the term "semantic network") and are provided by the European Interoperability Framework which governs data standards to ensure that data can be used across as many applications as possible.[3]

Next, these templates are used as extraction filters to transform unstructured human natural language (i.e. thousands of pages of raw text occurring on websites, online forms, etc.) into machine-readable semantic networks which can be utilised in software applications, such as chatbots. The process runs as follows: data is collected automatically from websites, then only those pages containing public-service information are selected. Then, paragraphs describing administrative proce-

[1] https://cefat4cities.eu/

[2] https://www.fiware.org/developers/catalogue/

[3] https://ec.europa.eu/isa2/solutions/core-public-service-vocabulary-application-profile-cpsv-ap_en



dures are extracted and syntactically analysed to identify nodes occurring in the template. Finally, relations between the extracted nodes are identified (the most challenging part of the process) and the information is delivered in a standardised linked open data (LOD) format compatible with the FIWARE Context Broker. Any follow-up effort or downstream software application can use this schema to subscribe to the created public-service content.

To achieve this, we resort to various multilingual Natural Language Processing (NLP) techniques such as automated classification, topic modelling, clustering, syntactic parsing and machine translation. When developing the solution, several challenging issues were identified. Discovering links between nodes (connecting for example an administrative procedure and all the evidence a citizen must provide to fulfil it) proved to be a non-trivial task. A unique solution had to be built, combining syntactic parsing and classification, since no out-of-the-box components existed to do this. Throughout the pipeline, a balance was needed between using monolingual and multilingual NLP models using translated data, since many linguistic NLP models only exist for a couple of languages. Finally, often the language itself was problematic. Current NLP models excel at "recognising" the meaning of a word when it appears within a larger body of text, but when words occur isolated (for example in a title or a table) recognition and translation become more difficult.

3 Outcome

The CEFAT4Cities project is currently coming to the end, but it has already impacted the way people think about public-service data in two major European Cities: The Brussels and Vienna Business agencies have successfully built a demonstrator chatbot with LOD generated by the CEFAT4Cities pipeline, and they realise that the data can be shared and used for other purposes.

Admittedly, the generated data still needs human validation, but considering the rate at which the CEFAT4Cities system outputs data and takes over the heavy lifting from humans (manually researching the business domain, clustering topics, creating the mental model, extracting intents, compiling and annotating the data sets, translating, etc.), there is plenty of time saved that can be used for fine-tuning the produced data sets.

The system currently exists as a prototype for the semantic modelling of public services in Croatian, Dutch, English, French, German, Italian, and Norwegian, with both the number of domains and languages expected to increase in the future.

From the onset of the project, the aim was to help smaller cities, as they have less means to build their own semantic network of public services, let alone to do this in a multilingual way. The extracted semantic network is abstract enough to allow for "knowledge transfer" between cities to build analogous systems. Looking at the first results, it is believed this ambition can be achieved, provided that a sufficient amount of evangelisation is carried out. Achieving this goal would greatly benefit smaller cities, as it will allow them to implement multilingual e-government solutions at a much faster pace and contribute to the free movement of EU citizens in general.

Acknowledgement

This project paper is an adapted and shortened version of a previously published blog post for the FIWARE Foundation (`https://www.fiware.org/2021/12/16/innovative-mind-mapping-system-connecting-to-smart-city-iot-networks/`). We would like to thank the EC's CEF Telecom programme for funding the CEFAT4Cities project (2019-EU-IA-0015) and the FIWARE Foundation for helping us in spreading the word.