Nuts4Nuts: extraction of geospatial information from Wikipedia and linking to OpenStreetMap

Nuts4Nuts uses a neural network to identify the municipality of the object of a Wikipedia article

Cristian Consonni
Digital Commons Lab, FBK-ICT, Fondazione Bruno kessler
consonni@fbk.eu — (+39) 0461 314 646
https://digitalcommonslab.fbk.eu/

Abstract

Volunteered geographical information (VGI) are one facet of phenomenon of crowdsourcing in which people are collecting and sharing large amounts data in open and collaborative projects. Although these projects have different purposes and scopes there is some overlap between them so it can be asked if these data, which are collected from different communities with different processes, are coherent.

In this context we have developed a tool, called *Nuts4Nuts*, which can identify the municipality in which a Wikipedia article is located extracting relevant informations from the templates or perfoming an analysis of the article's incipit.

The code is available with a permissive MIT license. At the moment, the system is limited to locations in Italy and is based on Italian Wikipedia.

Introduction

OpenStreetMap is a *free*, *editable*, map of the *whole world*, created online through volunteer effort[1]. The project allows the collection of features and its enrichment using tags attached to its basic data structures (nodes, ways, and relations). Each tag describes an attribute of the feature and may affect its rendering on the map. The tagging system is defined by the project's contributors and, whilst tags for common uses are described and reccommended, they can be defined and used freely provided their values are verifiable.

entries consist in text (or media) Wikipedia is a *multilingual*, *web-based*, *free-content* encyclopedia project and based on an openly editable model.

The two projects contain overlapping data (since they can refer to the same objects in te physical world) but the he processes used for collecting informations are different: whilst on Wikipedia anonymous users can edit, content is restricted only to encyclopedical subjects; whereas in OpenStreetMap registration is mandatory to edit the map and in principle every physical object (including extensively some non-physical information like administrative boundaries, bus route and similar) can be added to the map.

In Wikipedia entries consist in text and non-structured information while in OpenStreetMap entries consist in data.

Finally in Wikipedia content can be protected from editing in case of problems (called "vandalism", a term used also on OpenStreetMap), while in OpenStreetMap content is always editable.

In OpenStreetMap it is possible to link Wikipedia articles using a Wikipedia tags: $wikipedia{=}\langle language\rangle{:}\langle article\_title\rangle$. These tags provide a direct link between OpenStreetMap objects and Wikipedia articles. To date, to our knowledge, there is no way to link, within Wikipedia, OpenStreetMap objects from articles.

Main Objectives

Nuts4Nuts aims to provide a reliable and comprehensive tool for the recognition of the position of the object of a Wikipedia article.

The location returned as a result is a municipality (LAU2) or a sundivision of that (LAU3) in Italy.

Methods

The algorithm followed by Nuts4Nuts can be outlined as follows:

1. Analyze the template data
   (a) Extract address data;
   (b) Reconcile geographical entities; administrative unites with a known database;
2. Analyze the article's abstract
   (a) Get the article abstract and clean the formatting;
   (b) Perform entity recognition and select only geographical and place-like entities;
   (c) Perform entity reconciliation and selection as in steps 1.b and 1.c above;
   (d) Compute features for all the entities found;
   (e) Input data in a previously trained neural network;
3. Select the winning candidate and compute the final score
   (a) If template analysis (step 1 above) and neural network (step 2 above) identify common candidates return them with their score;
   (b) In the identification of the winning candidate an higher score is given to the candidates coming from template data;
   (c) Select only the LAU2 and LAU3s and form a set of candidates;
   (d) A candidate is decalared the "winner" if its score exceed a given threshold; Note that the

Implementation

Nuts4Nuts is written in Python and uses the PyBrain[2] library to define and train a simple feed-forward neural network and perform the abstract analysis (step 2 above). The layout of the neural network is depicted in fig. 1.

The entity extraction step has been performed using the *DataTXT-NEX* [3] and the reconciliation over administrative entites is performed using the reconciliation service *nutsrecon*[4] both provided by SpazioDati S.r.l. Nuts4Nuts has also been exposed as an *OpenRefine* reconciliation service, also provided by SpazioDati S.r.l., returning results in *JSON* format. The reconciliation service is available at the address:

http://nuts4nutsrecon.spaziodati.eu/.

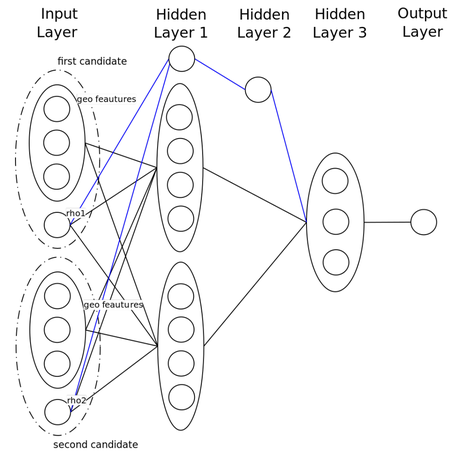

Figure 1: The layout of the neural network used in Nuts4Nuts

Known Limitations

Nunts4Nuts can find municipalities only in Italy, since it relies on the recognition of municipalities in Italy provided by *nutsrecon*, furthermore at the moment the entity extraction performed DataTXT-NEX is available only for text in Italian or English, in this light the scope of Nuts4Nuts has been limited to Italian Wikipedia.

Finally, it should be Nuts4Nuts makes no check to recognise if a request is sensible, i.e. if the article for which the location is requested can be placed on a map.

Results

A supervised learning approach has been taken to train the neural network of Nuts4Nuts. The training set has been created selecting manually 200 Wikipedia articles. To select only "mappable" articles, they were chosen randomly but kept only if they beloged to specific categories. A test set of the same size (200 samples) was created using the same procedure. The results of testing Nuts4Nuts are shown in 1.

Test Set #	Positive answer	Errors	No answer
1	185	3	12

Table 1: Perfomance of Nuts4Nuts over a test set of 200 Wikipedia articles in Italian Wikipedia.

Conclusions

- Nuts4Nuts can assign the municipality (in Italy) for an article from Italian Wikipedia using a feed-forward NN trained using supervised learning;
- Nuts4Nuts is shown to recognize the correct municipality (or a smaller administrative unit there contained) in the 92.5%.

Forthcoming Research

The development of Nuts4Nuts is inserted in a wider action aimed at the development of tools to enhance the work of the Italian and international communities of OpenStreetMap and Wikipedia. In particular forthcoming efforts will be focused on:

1. Compare the data
   (a) Identify links between Wikipedia pages and OSM entities
   (b) Extract all the available geographical information
   (c) Define metrics to calculate if the data are close or not
2. Reconcile the differences
   (a) Provide the communities with the result of previous analysis
   (b) Creating tools to facilitate the reconciliation

Nuts4Nuts as already been integrated in a tool developed by the Italian OpenStreetMap community (in particular by OpenStreetMap contributor *Simone F.*): *Wikipedia-tags-in-OSM*. *Wikipedia-tags-in-OSM* has been available online by Luca Delucchi (at Fondazione Edmund Mach, Trento) at the address:

http://geodati.fmach.it/gfoss_geodata/osm/wtosm/index.html

This project builds upon a similar one developed by the German Wikipedia community (and in particular Wikipedia contributor *Kolossos*): *WIWOSM*.

References

[1] M. F. Goodchild. "Citizens as sensors: the world of volunteered geography". In: *GeoJournal* 69 (2007), pp. 211–221. DOI: 10.1007/s10708-007-9111-y.

[2] Tom Schaul et al. "PyBrain". In: *J. Mach. Learn. Res.* 11 (Mar. 2010), pp. 743–746. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=1756006.1756030.

[3] SpazioDati s.r.l. *DataTXT-NEX (Named Entity eXtraction) service*. Jan. 2014. URL: https://dandelion.eu/products/datatxt/nex/demo/.

[4] SpazioDati s.r.l. *nutsrecon (NUTs reconciliation) service*. Jan. 2014. URL: http://nuts.spaziodati.eu/.

Acknowledgements

This work has been realized within the scope of the *T2DataExchange* project by *SpazioDati s.r.l.* and *Edizioni Curcu & Genovese* and funded from the *European Regional Development Fund*.