A semi-supervised Learning Approach to find equivalent long-string Organization Names

Frédérique Bordignon, Nicolas Turenne, Yann Feugueur

A semi-supervised Learning Approach to find equivalent long-string Organization Names

Frédérique BORDIGNON[1], Nicolas TURENNE[2], Yann FEUGUEUR[3]

[1]Direction de la documentation, Ecole des Ponts ParisTech, 6-8, av. Blaise Pascal, Cité Descartes - Champs-sur-Marne, F-77455 Marne la Vallée cedex 02
[2]Université Paris Est, LISIS, INRA, 77454 Marne-La-Vallée, France
[3]Orange, 39 Fbg de France, 90000 Belfort

correspondance: Frederique.bordignon@enpc.fr

Background

A platform called Opalia has been built to propose free access to all publications about a laboratory for a given range of years. This platform makes indexing of a corpus of scientific article of a given lab. But in the French research system a lab includes researchers from different organizations in a same unit generally called. UMR. Author can write differently their laboratory names.

Aim: Sorting a set of labels that is noisy can be seen as a binary classification into positives and leave negatives strings. We propose to use a cascade processing with help of tagging some positive strings to build a relevant space of features that helps classification into good labels.

Dataset constitution

Scopus®

THOMSON REUTERS WEB OF SCIENCE

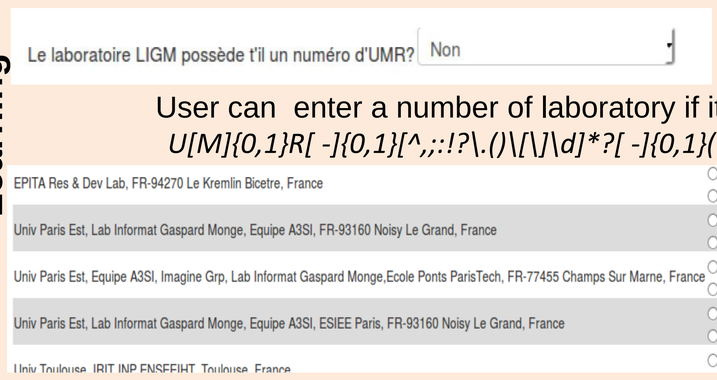
Data Acquisition procedure:

- ✓ From Scopus (http://www.scopus.com) or from the Web of Science (https://webofknowledge.com). Considered as the biggest ones, a gold reference and easy to grab.
- ✓ Query by keywords. For instance if a laboratory is called : *UMR 1326 LISIS, interdisciplinary research laboratory in science and innovations in societies, INRA, CNRS, ESIEE, UPEM*. The query will be “LISIS” in the field address field.

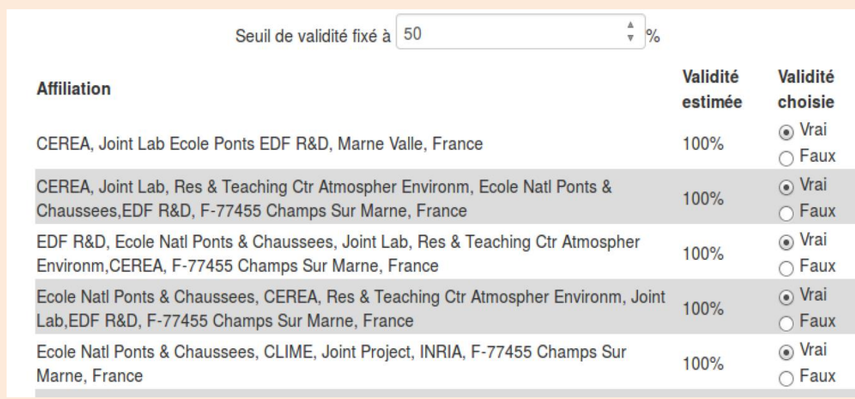
Main results and interpretations

User interface of the system. User upload its corpus (list of organization labels), name of address field; a laboratory label if learning has been already done.

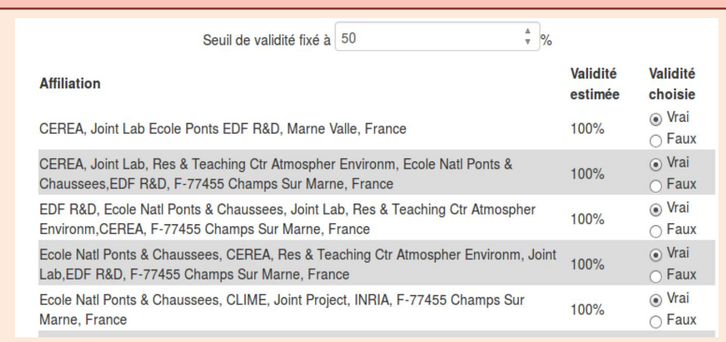
Learning



User can enter a number of laboratory if it exists
U[M]{0,1}R[-]{0,1}[^,;:!?\.()\[\]\d]?[-]{0,1}(\d{2,5})*



In other cases, User need to select positive examples (4 at least)



Configuration step → Feature acquisition step → Classification step

Algorithm 1 - feature Selection

1. Build a word (feature) histogram with wholeset of organization labels to process.
2. Grouping words by similarity.
3. Rewriting all Organization Labels (OL) with obtained features
4. Scan all Organization Labels to search a metadata as a laboratory number.
5. Scan all Organization Labels to build a list of valid features (require rewritten words)

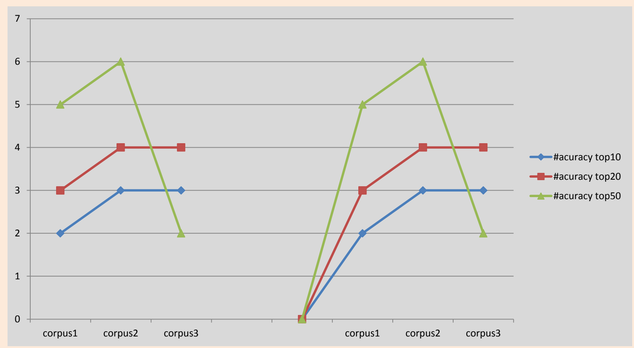
Algorithm 2 - Learning Model

1. Build a matrix taking into account OL selected by a user and metadata from previous steps. We have two kinds of OL, those coming from history of previous processing, and those coming from the given dataset to process. On the last type, we find the true positive OL (more than 80%) and true negative (less than 20%). These data make a comparison matrix.
2. Build a test matrix taking into account the rest of all others OL.
3. Set the two matrix as input to a SVM routine.
4. Assign to each OL of the test matrix, a relevance percent from the previous step and go back to step 1 till the list of words does not change any more.

Algorithm 3 - Classification Algorithm

1. Select valid OL to extract their word trigrams (reference dataset)
2. Build a matrix on a known OL (exact true/false one), each of these OL produce a vector. So parameter of KNN is k=1, because we have 2-class classification for OL (true or false).
3. Create a vector for each unknown OL, it will return a test matrix.
4. Compare the two matrices with Knn algorithm, each vector of test matrix will be clasified in the neighbouhood of the model matrix.
5. Assign to each OL of the test matrix, the obtained result, hence go back to step 1 till number of trigrams of the reference dataset does not change.

Classification Accuracy with a sample of 100 affiliations about 3 corpus



Perspectives

Local information, domain-dependant (as number of a laboratory) coupled to robust selection of lexical form leads to a good input for a learning step with a SVM model, and a classification step with KNN algorithm. The gain with such aided-computing interface a documentalist relies to easily classify similar address of laboratories which is time consuming. In perspective, the program will be used with several hundred of laboratories and possibly promote this service (https://exia.cortext.net/) in a Open-Access platform such as HAL (https://hal.archives-ouvertes.fr/).

Literature reference: [1] Bikel D.M., Schwartz R. and Weischedel R.M. An Algorithm that Learns What's in a Name, Machine Learning, February 1999, Volume 34, Issue 1, pp 211-231.

Acknowledgements: This work was funded by EXIA project (PEPS grant from Université Paris-Est and CNRS).