UNIOR NLP @ COFE Datathon 2022
Task 5 - Open task

Gennaro Nolano, Maria Pia di Buono, Johanna Monti
UNIOR NLP Research Group
Unversity of Naples ”L’Orientale”, Italy
{gnolano,mpdibuono,jmonti}@unior.it

Abstract

This paper presents the methodology proposed by the UNIOR NLP Research Group for Task 5 - Open Task of the COFE Datathon, a data mining competition organised by the European Commission in the context of the Conference on the Future of Europe (COFE).
We focus on taxonomy induction for a fine-grained classification of idea proposals.

1 Introduction

Within the COFE Datathon[1], the fifth challenge is an open task which means that participants are free to submit their own proposals to extract any type of information from the data provided by users and collected through the COFE digital platform.
Our team chooses to develop a workflow which aims at inducting a taxonomy, which provides a fine-grained classification of idea proposals thus improving the already existing COFE category information associated to proposals. Indeed, within the COFE dataset, proposals may present a category chosen by users to classify their contributions. Nevertheless, some idea proposals miss the user-generated classification. Furthermore, being user-generated, this classification may be inconsistent across proposals and users. For these reasons, we frame the task as taxonomy induction in order to improve the existing proposal classification.

2 Dataset

The dataset collected through the COFE digital platform includes: (i) about 17k idea proposals in several languages, (ii) an EN translation when needed, (iii) a user-generated category to classify the proposal, (iv) scope, (v) endorsement, (vi) anonymized users’ information, (vii) comments to the proposal. The provided dataset is characterized by several aspects that influence the workflow, those are:

- Multilingual data;
- Users’ generated data which may contain typos and a lack or misuse of domain terminology;
- Machine Translations in English (thus, susceptible to translation errors, mainly for under- represented languages)
- Lack of alignment between the source content and the English translation.
- Language classification of post content can be inaccurate, as it is not based on automatic language identification but on other users’ information. Thus, despite the provided language classification, some of the posts are actually written in English.
- Comments to the proposal without EN translation.

3 Methodology

The proposed workflow is based on three steps:

1. Keyword Extraction
2. Semantic Lexical Relation Extraction
3. Taxonomy Induction and Proposal Classification

Keyword Extraction This is is performed by means of the Python Keyphrase Extraction (PKE) module[2], which returns weighted results for the

[1] https://futureu.europa.eu/pages/datathon?locale=en

[2] https://github.com/boudinfl/pke

extracted keyphrases. Such results represent a first list of keywords which are then refined to include MultiWord Expressions (MWEs), according to the methodology we use for Task 1 of the COFE Datathon.

Semantic Lexical Relation Extraction After the keyword extraction phase, we use an external knowledge base, i.e., EuroVoc vocabularies[3], to extract semantic and lexical relation for each keyword. In order to in improve our results, we inject external knowledge from controlled sources, namely EuroVoc data, a multilingual and multidisciplinary thesaurus covering the activities of the EU and containing terms in 24 EU languages, available through a REST API service[4]. Entries in EuroVoc thesaurus are classified according to a taxonomy identifying hyperonims and hyponims (i.e., narrower and broader term), thus we use such information for the final phase of our workflow.

Taxonomy Induction and Proposal Classification EuroVoc hierarchical classification of concepts is used to include further conceptual information regarding proposals. In fact, while proposals are classified according to the COFE category they belong to, keywords might represent a source for additional data. This data can also be used with different granularity according to the level of hierarchy for EuroVoc concepts that is being taken into account, and its inclusion is beneficial to many tasks, e.g. topic detection.

4 Results and Conclusion

We present the workflow used to induct a taxonomy for classifying idea proposals.

The results of our workflow have been submitted to the COFE datathon and evaluate by the organizers. The final output is a tsv file in which each row has columns with the following information: the specific proposal id, the EuroVoc conceptual category for the proposal, and the EuroVoc conceptual categories for each of the top-5 highest scoring keywords.

As future work, we plan to evaluate our results against a baseline. We also intend to refine this classification, also integrating other external knowledge.

Furthermore, the integration of the full range of data available for the proposals (i.e., data regarding users, comments, etc.) would be useful in underlying new information about the data at hand, while also paving the way for other tasks such as social community detection and stance classification.

Acknowledgments

Maria Pia di Buono has been supported by Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 - Fondo Sociale Europeo, Azione I.2 “Attrazione e Mobilità Internazionale dei Ricercatori” Avviso D.D. n 407 del 27/02/2018.

[3] `https://op.europa.eu/en/web/eu-vocabularies`

[4] `https://www.vocabularyserver.com/eurovoc/`