UNIOR NLP @ COFE Datathon 2022
Task 1 - Keyword extraction from proposals

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 1 - Keyword Extraction 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 apply an approach based on the injection of external knowledge and a subsequent graph-based representation to select keywords from authomatically extracted candidates.

1 Introduction

Within the COFE Datathon[1], the first challenge aims at extracting keywords from conference open data, namely data provided by users and collected through the COFE digital platform.
Keyword extraction, namely the identification of the most relevant words or set of words which describe the central theme of any document (Abilhoa and De Castro, 2014; Lahiri et al., 2017; Mihalcea and Tarau, 2004), is one of the most elementary research for Natural Language Processing (NLP) and Information Retrieval (IR) (Garg, 2021).
The process of automatically extracting keywords is directly applicable to a wide range of NLP tasks, e.g., text summarization (Bharti and Babu, 2017; Lin, 2004; Litvak and Last, 2008), topic detection (Liu et al., 2010; Bougouin et al., 2013), event detection (Garg and Kumar, 2018).

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 a 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 four steps:

1. Automatic Keyphrase Extraction
2. MultiWord Expression (MWE) Discovery
3. External Knowledge Injection
4. Graph-based Candidate Representation

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

Automatic Keyphrase Extraction This step is performed by means of the Python Keyphrase Extraction (PKE) module[2], which returns weighted results for the extracted keyphrases. In particular, we opt for the MultiPartite Ranking (Boudin, 2018) algorithm to extract the keywords. This algorithm both extracts keywords while ranking them according to their relevancy to the text by giving them a score between 0 and 1.

Such results represent a first list of scored keyword candidates which are used as inputs for the following workflow steps.

MultiWord Expression Discovery This step aims at discovering MWEs within the text. After the keyphrase extraction phase, we check the extracted candidates in context, i.e., sentence/proposal, and assume the candidate is part of a MWE when we encounter two or more candidates (w_n) close to each other or separated by specific elements, according to some hand-defined rules.

We set the maximum window size between two candidates to three elements and use Part of Speech (PoS) information to select sequences of words that can be MWE candidates according to some pre-defined patterns of co-occurring elements (Table 1). We check each candidate occuring either in the w_1 or w_2 position.

To assign a score to the extracted MWEs, we calculate the average MultiPartite Ranking value for each of the keywords involved in the MWE by summing the values for each keyword, and then dividing the result by the number of keywords beloging to the MWE.

Distance	Pattern
Zero	w_1 w_2
1 Element	w_1 *Preposition* w_2
	w_1 *Adjective* w_2
2 Elements	w_1 *Preposition Determiner* w_2
	w_1 *Adjective Preposition* w_2
3 Elements	w_1 *Adjective Preposition Determiner* w_2

Table 1: MWE candidate patterns.

External Knowledge Injection In order to improve the results from the previous step, we inject external knowledge from controlled sources, namely EuroVoc vocabularies[3].

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

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

EuroVoc data comprise of a multilingual and multidisciplinary thesaurus covering the activities of the EU and containing terms in 24 EU languages. For this particular work, we access the information from EuroVoc through a REST API service[4].

Entries in EuroVoc thesaurus are classified according to a taxonomy identifying hyperonyms and hyponyms (i.e., narrower and broader term).
In order use information from EuroVoc to filter out our results, we first perform a manual mapping between COFE data categories (from che `category/name/en` field in the dataset) and EuroVoc upper-class concept classification. Through this we extract, for each proposal, the corresponding EuroVoc concept for its category. Then, through the API, we look for the top-3 closest matches for every keyword, and we store for each of these match their corresponding hierarchical structure. For languages not supported by the API, we refer to the keywords extracted from the English translation.

With these data, we build a graph that is then used to filter the final keywords.

Graph-Based Candidate Representation For each proposal, we create a directed graph through the following steps:

1. a root node is created for the proposal, which is then linked to a target node representing its COFE category (if available);
2. the COFE category node, if present, is then linked to a node representing its mapped EuroVoc concept;
3. for each candidate keyword, a node is created, connected to the proposal's root node;
4. since keywords might be ambiguous, each of them is connected to the top-3 closest categories from EuroVoc;
5. each EuroVoc's category is connected to its broader representation, until the top concept is reached.

While generally the links of this graph have a default weight of 1, certain features are used to give more weight to specific features:

[4] https://www.vocabularyserver.com/eurovoc/
As of the time of writing this report, the API supports only 12 of these languages (Bulgarian, Spanish, Czech, German, Greek, English, French, Italian, Dutch, Polish, Portugues and Slovenian).

- the edge from keywords to their corresponding proposal's node has weight = 1 if the keyword is present in the title, weight = 0.3 otherwise;
- the edge from each keyword to its closest concept category from EuroVoc is 1, and each following edge up to the top concept has a weight equal to $1/log(d)$, where d is the number of steps dividing the object node from the keyword's node.

Such graph, exemplified in Figure 1, is used to re-rank the candidate keywords on the basis of their correlation to topics and concepts. In particular, for each keyword's node we calculate its betweenness centrality (Freeman, 1977), which is integrated together with the original results from the MultiPartite Rank algorithm.
After several tests, we opted to use the value $v = e^{mpr_{kw}} + e^{bc_{kw}}$, where mpr_{kw} is the value from the Multipartite Ranking, and bc_{kw} the betweenness centrality of a specific keyword.

4 Results and Conclusion

We present the workflow used to extract keywords from idea proposals, with a particular focus on MWE extraction.
The results of our workflow have been submitted to the COFE datathon in the form of tsv files, with each column containing the ID of the proposal, followed by the top-5 keywords extracted. The output thus formed was evaluated by the organizers.
As future work, we plan to evaluate our results against a baseline. We also intend to refine this extraction defining more rules to identify MWEs (e.g., by filtering which prepositions can be used for specific MWE patters), while also integrating other external knowledge. One of the main objectives would be to make the system able to deal with more languages.
We also plan on investigating the effects of other algorithms for authomatic keyword extraction (e.g., YAKE) and other measures for node centrality (e.g. PageRank).

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.

References

Willyan D Abilhoa and Leandro N De Castro. 2014. A keyword extraction method from twitter messages represented as graphs. *Applied Mathematics and Computation*, 240:308–325.

Santosh Kumar Bharti and Korra Sathya Babu. 2017. Automatic keyword extraction for text summarization: A survey. *arXiv preprint arXiv:1704.03242*.

Florian Boudin. 2018. Unsupervised keyphrase extraction with multipartite graphs.

Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. Topicrank: Graph-based topic ranking for keyphrase extraction. In *International joint conference on natural language processing (IJCNLP)*, pages 543–551.

Linton Freeman. 1977. A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41, 03.

Muskan Garg and Mukesh Kumar. 2018. The structure of word co-occurrence network for microblogs. *Physica A: Statistical Mechanics and its Applications*, 512:698–720.

Muskan Garg. 2021. A survey on different dimensions for graphical keyword extraction techniques. *Artificial Intelligence Review*, 54(6):4731–4770.

Shibamouli Lahiri, Rada Mihalcea, and P-H Lai. 2017. Keyword extraction from emails. *Natural Language Engineering*, 23(2):295–317.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Marina Litvak and Mark Last. 2008. Graph-based keyword extraction for single-document summarization. In *Coling 2008: Proceedings of the workshop multi-source multilingual information extraction and summarization*, pages 17–24.

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 366–376.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

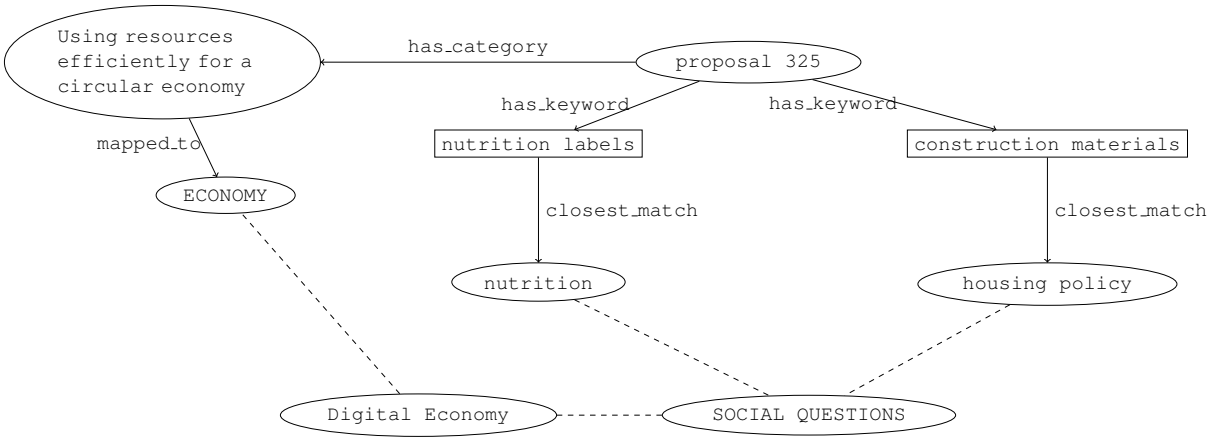

Figure 1: Example of a graph used for keyword extraction. Dashed edges indicate the presence of a path connecting the two nodes.