From Monolingual Multiword Expression Discovery to Multilingual Concept Enrichment: an Ontology-based approach

Gennaro Nolano[1], Maria Pia di Buono[1], and Johanna Monti[1]

[1]Unior NLP Research Group, University of Naples "L'Orientale"
{gnolano,mpdibuono,jmonti}@unior.it

Abstract. In this paper, we present a methodology for the semantic enrichment of cultural heritage (CH) data, based on the use of ontologies and Linked data. The proposed method aims at developing domain-specific resources enriched with multilingual conceptual information starting from monolingual RDF data. Particularly, our approach begins with a Multiword Expressions (MWEs) discovery process to select a starting list of domain-specific candidate mentions. Subsequently, we perform a concept discovery phase in order to link them to closely matching Dbpedia concepts through the use of two similarity measures. The semantic information related to these concepts is used to further filter the candidates and obtain representative mention-concept pairs by reweighting automatically computed scores making use of a graph representation.
We test our methodology on biographic information about authors extracted from the Europeana Data Collection. The final results are a resource of semantically enriched data, containing a list of domain-specific keywords and MWEs together with Dbpedia concepts they strongly match, and the multilingual labels representing these specific concepts.

Keywords: MWE discovery · Concept Discovery · Ontology

1 Introduction

In this paper, we present an approach to developing domain-specific multilingual resources starting from monolingual RDF data. Particularly, our approach focuses on Multiword Expressions (MWEs) discovery and exploits linking techniques through similarity measures to enrich the final linguistic resource (LR). The main rationale behind the proposed methodology is that RDF metadata contains fields for both descriptive texts (e.g., `dbo:abstract`[1]) and structured information from external Knowledge Bases (KBs) (e.g., `dbo:wikiPageExternalLink`[2]). These two sources of data could both be exploited in the creation of a domain-specific semantically enriched resource of entities and mentions (i.e., entities'

[1] `https://es.dbpedia.org/ontology/abstract`
[2] `https://dbpedia.org/ontology/wikiPageExternalLink`

Proceedings of the International Conference EUROPHRAS 2022, pages 198–207,
Malaga, Spain, 28-30 September, 2022.
https://doi.org/10.26615/978-954-452-080-9_024

surface forms) from descriptive texts are extracted and conceptually linked to structured information. A resource created in this way would be suitable for a series of Natural Language Processing (NLP) tasks, such as Terminology Extraction, Machine Translation and Entity Linking.
The paper is organized as follows: Section 2 describes some of the efforts made by researchers in the fields of MWE discovery and Semantic Enrichment; Section 3 explains in general terms the proposed methodology; Section 4 is devoted to the practical aspects related to the creation of the proposed LR; finally, Section 5 illustrates the final LR while also introducing possible future work.

2 Related Work

Multiword Expression Discovery There is a considerable body of works describing techniques to automatically detect MEWs. Generally, this is solved through statistical means by computing the correlation strengths between words forming the expression [14,6], with the most widely used association measure being pointwise mutual information [5]. Despite the effectiveness of these models, one of the main drawbacks is the need for a background corpus to compute statistical significance. This corpus might not exist, or might not be big enough when dealing with certain domains and certain languages.
Another option is to use syntactic patterns to generate MWE candidates. This, for instance, has been explored in works such as [11,1]. Since these patterns can be generated from heuristics, they can be applied to any kind of text, no matter their length. In this work, we propose the use of several syntactic patterns specifically tailored for the Italian language.

Semantic Enrichment Much effort has been made trying to fill the semantic gap between the "web of documents" and the "web of knowledge" [4], as shown in works such as [7,2,8,16].
Nevertheless, such models have generally focused on tasks related to Named Entities, such as Named Entity Recognition (NER), Named Entity Linking (NEL) and Named Entity Disambiguation (NED). While these tasks effectively integrate some sort of semantic knowledge into raw texts, they generally focus on just Named Entities which would directly leave out important domain-specific concepts, such as classes and topics (e.g., *classical music* and *Roman architecture*), which are rarely identified by proper names.
For this reason, connecting important spans of text to *concepts* rather than specific Named Entities can benefit many NLP tasks. Concept discovery [12] has been explored for domains such as news articles [10] and scientific knowledge [15].
One drawback of such models is that they generally focus on a single language, while connecting a raw text to specific concepts present in a knowledge base as Wikidata or Dbpedia can help provide multilingual access to data, thus improving the reusability of LR.
In this work, we integrate multilingual data in the final resource by exploiting labels used to describe Dbpedia concepts.

3 Methodology

Our methodology relies on the use of monolingual RDF data and makes use of the unstructured text presented by descriptive metadata and structured information in the form of external links. Basically, we extract MWEs from the unstructured texts and connect them to the concepts (in the form of links) which are conceptually related to the specific RDF item that is being described. More specifically, we perform three main steps:

1. Monolingual MWE Discovery
2. Concept Discovery
3. Ontology-based filtering

The first two steps of the methodology are shown in graphical form in Figure 1, while an example of a graph used for ontology-based filtering is shown in Figure 2.

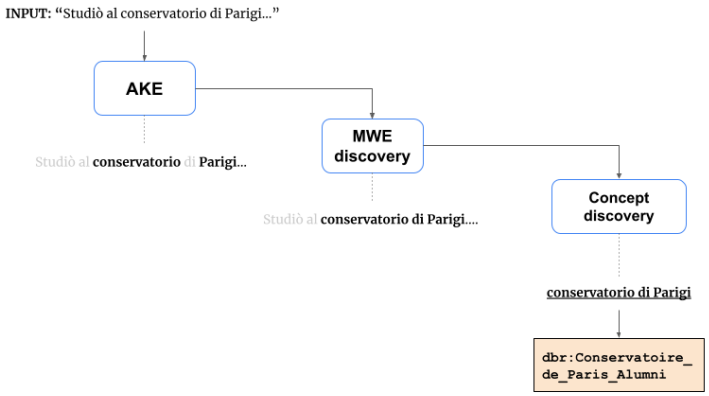

Fig. 1. Graphical representation of the process of monolingual mwe discovery: in the first step, Keywords are automatically extracted using the pke Python library. The automatically extracted keywords are then expanded based on specific patterns. The extracted MWEs are connected to the link they most likely refer to, through similarity measure with the italian label of said link.

Monolingual MWE Discovery In order to perform this step, we first rely on off-the-shelf libraries to extract keywords from the texts at hand. While some of these keywords might be represented by Named Entities, we do not want to put any restriction on the type of semantic information we want to extract from the

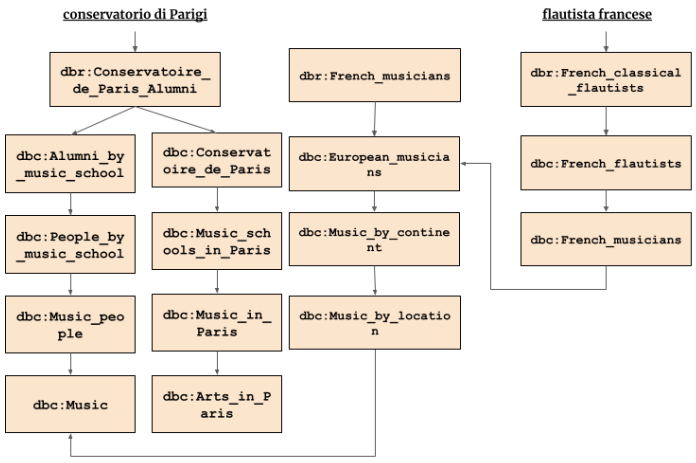

Fig. 2. Graphical representation of the graph implemented for ontology-based filtering. The graph is constructed traversing the `skos:broader` property 4 times from the links available for the entity. The nodes representing concepts related to the entity will generally share several edges, thus giving more weights to the MWEs linked to them.

texts. Thus, we do not rely on NER tools (which usually limit their information to People, Locations and Organizations), but rather automatic keyword extraction tools, which rely on purely statistical information. The extracted results can thus be any sort of concept, without restrictions.
These results represent a first list of scored keyword candidates which are then used as inputs for an MWE discovery phase.
To achieve that, after the automatic keyphrase extraction phase, we check the extracted candidates in context, that is in the source texts, and assume that a candidate w_x is part of an MWE when it is close to another candidate w_y and they are separated by specific elements, according to hand-defined linguistic patterns based on heuristics (Table 1).
Then, we assign a relevance score to the extracted MWEs and classify them using the conceptual category already associated with the main item.

Concept Discovery In order to improve the results from the previous step and semantically enrich them, we inject external knowledge from DBpedia[3]. Since most RDF data are linked to other external KBs, it is possible to include it as a source of additional information, which can usually be accessed through specific endpoints. In particular, the data used in this work is connected to DBpedia entries, and we make use of the DBpedia SPARQL endpoint[4] and the `dbo:wikiPageWikiLink` property, which connects a specific DBpedia entry to

[3] `https://www.dbpedia.org/`
[4] `https://dbpedia.org/sparql`

Distance	Pattern	Example
Zero	w_1 w_2	Adelaide Festival
1 Element	w_1 *PREP* w_2	nuova generazione *di* musicisti folk
	w_1 *ADJ* w_2	Aleksandra Aleksandrovna *nata* Grigorovič
2 Elements	w_1 *PREP DET* w_2	premio Nobel *per la* letteratura
	w_1 *ADJ PREP* w_2	direttore *principale della* Philharmonia Orchestra

Table 1. MWE candidate patterns.

The translation for the presented examples are, respectively: *Adelaide Festival*, *new generation of folk musicians*, *Aleksandra Aleksadrovna nee Grigorovič*, *Nobel prize for Literature*, *main directory of the Philharmonic Orchestra*.

every other entry present as a link in its Wikipedia article. This means that concepts that are related to a certain topic or domain are likely to share similar links.
Once these entities are extracted, we calculate similarities between them and extracted MWEs. Thus, it is possible to create a network of connections between raw spans of text and entities in DBpedia, consequently obtaining a certain level of semantic enrichment, while also enabling the inclusion of multilingual data in the form of the labels representing the various concepts.

Ontology-based Filtering The information collected during the concept discovery phase is then used to filter the extracted candidates. All the links extracted in the previous steps are also used to recreate a hierarchical graph by traversing the `skos:broader` relation. In particular, for each link, we extract its parent nodes until we get to the 4rd highest node in the hierarchy.
This way, we end up with a graph-based representation of the connected concepts, in particular the links that were connected to MWEs during concept discovery. We make use of this graph to filter and reweight domain-specific keywords and MWEs, while also using it as a source of additional semantic information to enrich the data at hand.

4 Experiment and Results

Data Collection In order to collect domain-specific texts, we refer to the Europeana Entity API[5], which allows for the search and retrieval of RDF data about entities from the Europeana Entity Collection.
These entities represent a collection of Named Entities harvested from and linked to several online data catalogues, such as Geonames, DBpedia and Wikidata. In particular, for the purpose of this work, we extract biographical information about entities of type `agents`, which represent artists from different cultural heritage sub-domains such as music and fashion.

[5] `https://pro.europeana.eu/page/entity`

Using the SPARQL API for the Europeana Data Collection[6], we recollect the following information for 500 agent entities:

- their English label,
- the DBpedia entry they are linked to, and
- the Italian text for their biographical information.

Monolingual MWE discovery We first extract keyphrases from each text using the `pke` Python library[7], which returns a list of weighted results representing extracted keyphrases and their relevance according to a specific model. The library provides several models for AKE, among which we opt for the MultiPartite Ranking [3]. The resulting relevance score for each keywords ranges from 0 to 1. As already stated, from this list of automatically extracted keywords, we aim at discovering MWEs within the text. To do so, we check whether two keyphrases are close enough[8], and whether the sequence of words between them is acceptable according to pre-defined patterns of co-occurring elements (as shown in Table 1). We check each candidate occurring either in the w_1 or w_2 position.
To assign a score to the newly extracted MWEs, we calculate the average MultiPartite Ranking value for each of the keyphrases involved in the MWE by summing the values of each keyphrase and then dividing the result by the number of keyphrases belonging to the discovered MWE. Keyphrases which cannot be used to build any new MWE are kept as they are.
In total, we extract 4770 keywords and MWEs. In Table 2 we show the different effectiveness of each linguistic pattern in discovering new MWEs, together with the number of linked concepts for each of these patterns, as described in the next paragraph.

Pattern	**Occurrences**	**Linked to Concepts**
w_n	3795	1620
w_1 w_2	12	6
w_1 PREP w_2	878	555
w_1 ADJ w_2	3	1
w_1 PREP DET w_2	66	38
w_1 ADJ PREP w_2	16	12
Total	**4770**	**2232**

Table 2. Number of MWEs and connected concepts

Concept Discovery For each agent entity, we exploit its entry in DBpedia to extract all the hyperlinks present in the entity's corresponding Wikipedia page by accessing the `dbo:wikipageWikiLink` property using the DBpedia SPARQL

[6] `http://sparql.europeana.eu/`
[7] `https://github.com/boudinfl/pke`
[8] In this work we set the maximum distance window at 2 tokens.

endpoint.
One of the main issues of such hyperlinks is that in some cases they only present labels for the English language. This is the case for most of the category-defining entries such as `dbp:Victorian_poets` and `dbc:19th-century_English_poets`. Since these links generally refer to domain-specific knowledge classification, we want to access them even in absence of an Italian label. In order to do so, in case an Italian label is unavailable for a specific link, we automatically translate it from the English label using the Argos Translate Python library[9]. In order to make full use of the linked information available on DBpedia, we connect each keyword and MWE to the concept it most closely matches, by applying similarity measures over the links present in each specific page.
In particular, we use pre-trained fastText word vectors for Italian[10] to represent both MWEs and the Italian labels in vector space. Once we obtain these distributional representations, for each agent entity we compute the similarity scores between each MWE and each page link's Italian label. The similarity scores are calculated as the raw product of two different measures: cosine similarity between the embeddings and the overlap coefficient (i.e., the Szymkiewicz–Simpson coefficient [13]) between the surface forms. This way, we take into account semantic similarity between vectors, while also accounting for cases in which a specific MWE and a label share similar surface form despite their vectors being distant. From this list of computed similarity scores, we discard any MWE-link pair with a score lower than 0.4. Then, for each remaining MWE, we keep the link with the highest similarity score as the closest match. In case a MWE is not linked to any concepts, we leave it as it is for the following steps.
Regarding the 2232 keyphrases linked to Dbpedia concepts, in Table 3 we report the number of translated labels for each available languages in the final resource.

Language	**#Concepts**	**Language**	**#Concepts**
ar	933	ja	1028
ca	1.005	ko	893
cs	948	nl	1.085
de	1.228	pt	1.075
el	746	pl	1.075
en	2.232	ru	1.131
eu	815	sv	1.054
fr	1.266	uk	1.007
ga	514	zh	948
in	823		
Total			**19.806**

Table 3. Number of translated concepts

[9] `https://pypi.org/project/argostranslate/`
[10] `https://fasttext.cc/docs/en/crawl-vectors.html`

Ontology-based filtering Starting from the links collected from the entity, we recreate a hierarchical graph by traversing the `skos:broader` property 4 times. In this graph, concepts that are related to each other will generally share edges connecting to common nodes. In particular, we are interested in the links connected to specific MWEs.
We can then use the graph we obtain to re-rank the candidate keywords on the basis of their correlation to topics and concept: for each node linked to a specific MWE we calculate its betweenness centrality [9], which is then integrated together with the score calculated in the previous step.
For each MWE, the final score is computed as the sum $v_final = v_{mwe} + bc_{node}$, where $node_{mwe}$ is the value of the specific MWE, and bc_{node} the betweenness centrality of a the node connected to it. In case a MWE is not connected to any link, its score will be left as it was originally.
Finally, the extracted MWEs, ranked according to their reweighted scores, are enriched with related concepts and their multilingual labels (when present) from DBpedia.

5 Conclusion and Future Work

In this paper we described the process of semantic enrichment employed in the creation of a domain-specific multilingual resource.
The development makes use of statistical information to extract keyphrases, linguistic patterns to discover new MWE on the basis of automatically extracted keyphrases, similarity measures to link those to close concepts in Dbpedia, and finally graph-based representations to combine all these information together. The final proposed resource[11] is a collection of the following data extracted from the original biographic texts for 500 agent entities:

- MWE discovered through the combination of AKE and linguistic patterns,
- Dbpedia correlated entities linked to similar MWEs,
- multilingual labels for the Dbpedia entities for all available languages,
- broader concepts for each of the Dbpedia entity linked to a specific MWE.

In future work, we aim at improving the current results by refining the current process. For instance, a domain-specific BERT-like embedding model might help improving the concept discovery stage. The MWE discovery stage: for instance would benefit from a more flexibile and lexically (rather than just sintactically) grounded set of linguistic patterns might help improving current results while also reducing the noise currently present in the resource.

Acknowledgements

Maria Pia di Buono has been supported by Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 - Fondo Sociale Europeo, Azione I.2 “Attrazione

[11] `https://github.com/Nolanogenn/multilingual_ontology_based_ch`

e Mobilità Internazionale dei Ricercatori" Avviso D.D. n 407 del 27/02/2018. Authorship Attribution is as follows: Gennaro Nolano is author of Section 2 and Section 4, Maria Pia di Buono is author of Section 1 and Section 3, and Johanna Monti is author of Section 5 and supervised the project.

References

1. Baldwin, T.: Deep lexical acquisition of verb–particle constructions. Computer Speech Language **19**(4), 398–414 (2005). https://doi.org/https://doi.org/10.1016/j.csl.2005.02.004, https://www.sciencedirect.com/science/article/pii/S0885230805000070, special issue on Multiword Expression
2. Batchelor, C.R., Corbett, P.T.: Semantic enrichment of journal articles using chemical named entity recognition. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. pp. 45–48. Association for Computational Linguistics, Prague, Czech Republic (Jun 2007), https://aclanthology.org/P07-2012
3. Boudin, F.: Unsupervised keyphrase extraction with multipartite graphs (2018). https://doi.org/10.48550/ARXIV.1803.08721, https://arxiv.org/abs/1803.08721
4. Buitelaar, P., Cimiano, P.: Bridging the gap between text and knowledge **167**, v–ix (01 2008)
5. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics **16**(1), 22–29 (1990), https://aclanthology.org/J90-1003
6. Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., Todirascu, A.: Survey: Multiword expression processing: A Survey. Computational Linguistics **43**(4), 837–892 (Dec 2017). https://doi.org/10.1162/COLI_a_00302, https://aclanthology.org/J17-4005
7. Desmontils, E., Jacquin, C., Simon, L.: Ontology enrichment and indexing process (06 2003)
8. Dojchinovski, M., Sasaki, F., Gornostaja, T., Hellmann, S., Mannens, E., Salliau, F., Osella, M., Ritchie, P., Stoitsis, G., Koidl, K., Ackermann, M., Chakraborty, N.: FREME: Multilingual semantic enrichment with linked data and language technologies. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 4180–4183. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://aclanthology.org/L16-1660
9. Freeman, L.: A set of measures of centrality based on betweenness. Sociometry **40**, 35–41 (03 1977). https://doi.org/10.2307/3033543
10. Hassanzadeh, O., Trewin, S., Gliozzo, A.: Semantic Concept Discovery over Event Databases, pp. 288–303 (06 2018). https://doi.org/10.1007/978-3-319-93417-4_19
11. Justeson, J.S., Katz, S.M.: Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering **1**(1), 9–27 (1995). https://doi.org/10.1017/S1351324900000048
12. Lin, D., Pantel, P.: Concept discovery from text. In: COLING 2002: The 19th International Conference on Computational Linguistics (2002), https://aclanthology.org/C02-1144
13. M.K, V., Kavitha, K.: A survey on similarity measures in text mining (2016)

14. Pecina, P., Schlesinger, P.: Combining association measures for collocation extraction. In: Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. pp. 651–658. Association for Computational Linguistics, Sydney, Australia (Jul 2006), https://aclanthology.org/P06-2084
15. Shen, Z., Wu, C.H., Ma, L., Chen, C.P., Wang, K.: SciConceptMiner: A system for large-scale scientific concept discovery. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. pp. 48–54. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-demo.6, https://aclanthology.org/2021.acl-demo.6
16. Smrz, P., Otrusina, L.: Semantic enrichment across language: A case study of Czech bibliographic databases. In: Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017). pp. 523–532. NLP Association of India, Kolkata, India (Dec 2017), https://aclanthology.org/W17-7563