OKgraph: Unsupervised Structured Data Extraction from Plain Text (Extended Abstract)

Maurizio Atzori*
atzori@unica.it
Department of Math/CS
University of Cagliari (Italy)

Simone Balloccu
s.balloccu@studenti.unica.it
Department of Math/CS
University of Cagliari (Italy)

Andrea Bellanti
a.bellanti@studenti.unica.it
Department of Math/CS
University of Cagliari (Italy)

Emanuele Mameli
e.mameli@studenti.unica.it
Department of Math/CS
University of Cagliari (Italy)

Stefano Raimondo Usai
s.usai16@studenti.unica.it
Department of Math/CS
University of Cagliari (Italy)

ABSTRACT

In this report we introduce *OKgraph*, a software library for (open) Knowledge Graph extraction from free text. Named after a two-year project where we studied and developed unsupervised algorithms addressing tasks related to taxonomy learning, the library contains NLP tools powered by these results.

KEYWORDS

word embeddings, knowledge graphs, unsupervised learning, machine understanding

1 INTRODUCTION

As introduced in [1], *OKgraph* has been a two-year project funded by Regione Autonoma della Sardegna focused on investigating the fundamental relationship between unstructured data (natural language text) and structured data (graphs representing knowledge), eventually leading to an autonomous way of inferring the latter from the former. In this short abstract we introduce the main outcome of our research, a library named after the project that learns meaningful graph triples autonomously from scratch, that is, from non-annotated free text such as Wikipedia. We followed a statistical approach, focusing on a number of subtasks described next.

2 THE LIBRARY

OKgraph is a python3 library that performs unsupervised natural-language understanding (NLU). It currently addresses the following tasks:

- **Set Expansion** (or *co-hyponyms discovery*): given one or a short set of words, continues this set with a list of other "same-type" words (co-hyponyms)
- **Relation Expansion**: given one or a short set of word pairs, continues this set with a list of pairs having the same implicit relation of the given pairs
- **Set Labeling** (or *hypernym discovery*): given one or a short set of words, returns a list of short strings (labels) describing the given set (its type or hypernym)
- **Relation Labeling**: given one or a short set of word pairs, returns a list of short strings (labels) describing the relation in the given set.

Being unsupervised, it only takes a free (untagged) text corpus as input, in any space-separated language. Scriptio-continua corpora and languages needs third-party tokenization techniques.

3 SET EXPANSION

Given a small set of words it will expand it with "same type" words:

$$\{Italy, France, Germany\}$$
$$\downarrow$$
$$\{Spain, Denmark, Belgium...\}$$

In terms of graphs and taxonomy, the set expansion task is intended to get co-hyponyms (i.e., node siblings) from an initial set. It is crucial in graph construction, as it basically provides list of same-type nodes. We focused on singleton expansion [2], that expanding a set with cardinality 1, exploiting word embeddings similarity, and transitivity and symmetry of the same-as relation, that is:

(1) **Simmetric**: given two co-hyponyms, their vectors should be simmetrically near.
(2) **Transitive**: given three co-hyponyms transitivity should hold. For example given (*Italy*, *Germany*, *Spain*) if *Italy* is near *Germany* and *Spain*, then *Germany* must be near *Spain*. On the opposite given (*Italy*, *Rome*, *Germany*) we find that *Rome* breaks the bond with *Germany*.

The Set Expansion algorithms were evaluated by using 10-words sets belonging to a fixed number of categories; the testset was manually tagged and we obtained a significant improvement w.r.t. standard word2vec similarity:

Method	P@5	P@25	P@50
NNS	0.73	0.59	0.53
DEPTH	0.70	0.66	0.61
2WC	0.86	-	-
T5M	0.85	0.68	0.60

At the time of writing *OKgraph* is the only library which is known to perform singleton expansion.

*contact author. Supported in part by Sardegna Ricerche (CRP 120) and MIUR PRIN project *HOPE - High quality Open data Publishing and Enrichment.*

IIR 2019, September 16–18, 2019, Padova, Italy

We are also studying the problem of finding optimal vectors for hypernyms whose neighbours represents all hyponyms, using the powell optimization method with promising results.

4 RELATION EXPANSION

Given one or a set of word pairs will expand it with a list of pairs having the same implicit relation:

$$\{(Italy, Rome), (Germany, Berlin)\}$$
$$\downarrow$$
$$\{(Spain, Madrid), (France, Paris)...\}$$

This task is needed to extract edges connecting nodes in the final knowledge graph. We exploit set expansion results over the two different dimensions of the pairs, and also linearity of word embeddings, computing the centroid of $(Rome - Italy)$ and $(Berlin - Germany)$, representing the vector direction of the edge "being capital of".

5 SET LABELING (HYPERNYM DISCOVERY)

The Set Labeling task is currently under testing and will be introduced in the library shortly. Given one or a set of words, returns a list of short strings (labels) describing the given set:

$$\{apple, pear, pineapple\} \rightarrow \{fruit, food...\} \quad (1)$$

This task allows the system to autonomously learn the taxonomy from the text. This translates into a possible enhancement of the other tasks. Again, we exploit word embeddings with the following heuristics:

(1) Given a certain set S of words, and a set of potential labels, the hypernyms should have a very low variance, in terms of cosine similarity, with S.
(2) Hypernyms should also be more frequent than co-hyponyms.

We evaluated our unsupervised approach using the SemEval 2018 Task 9[3] benchmark, outperforming all others unsupervised algorithms and some supervised ones on all of the general-purpose corpora (1A, 1B, 1C). The following table shows the evaluation results on 1A corpus for the given task:

Method	superv.	P@5	P@15
CRIM r1	✓	19.03	18.27
CRIM r2	✓	18.74	17.98
MSCG-SANITY r1	✓	11.60	10.28
vanillaTaxoEmbed (baseline)	✓	9.91	9.26
NLP HZ	✓	9.19	8.78
MSCG-SANITY r2	✓	9.74	8.46
300-sparsians r1	✓	8.63	8.13
OKgraph Set Labeling		**4.70**	**8.07**
300-sparsians r2	✓	8.23	7.91
MFH (baseline)	✓	7.81	7.53
SJTU BCMI	✓	5.96	5.78
Team 13		2.72	2.48
Apollo r2		2.69	2.42
apsyn (baseline) r1		1.39	1.34
balapinc (baseline) r1		1.30	1.30
Apollo r1		1.33	1.22
slqs (baseline)		0.61	0.61

6 RELATION LABELING

Given one or a short set of word pairs, returns a list of short strings (labels) describing the relation in the given set:

$$\{(Italy, Rome), (France, Paris)\}$$
$$\downarrow$$
$$\{(capital, soccer\ team...)\}$$

This task is important to automatically label edges in a graph. The desired relationships cannot be obtained by the use of classic means provided by the NLP and a classical text analysis would lead to a high number of false positives. To obtain an effective Relation Labeling algorithm, different statistical and linguistic knowledge have been used:

(1) **Relative frequency**: given a sentence that contains the two entities of interest ("window"), the whole text frequency serves stopwords removal, and windows frequency helps labels discovery.
(2) **Windows parameterization**: too large windows introduces noise, the opposite situation lead to information loss.
(3) **Window dynamisation**: adoption of a variable distance between windows leads to better results.

The Relation Labeling algorithm works both on single and multiple pairs, through a logarithmic ranking that made it possible to evaluate the system on a manually tagged testset.
In order to evaluate the algorithm performance we used a set of 20 pairs for a fixed number of categories and evaluated the P@K for each one. The following table shows our evaluation results:

Category	P@1	P@5	P@10	P@15	P@20
cities/states	35.00	65.00	80.00	85.00	90.00
soccer player/teams	0.00	30.00	30.00	40.00	50.00
currencies/countries	10.00	25.00	75.00	85.00	85.00
founders/company	10.00	65.00	70.00	70.00	70.00
singers/groups	0.00	60.00	80.00	90.00	90.00
Total	11.0	49.00	76.00	74.00	77.00

7 CONCLUSIONS

We introduced OKgraph, a set of tools developed for Open Knowledge Graph extraction from free (plain) text. Our opensource software library will be available for download at https://github.com/okgraph

REFERENCES

[1] Maurizio Atzori. 2017. The Need of Structured Data: Introducing the OKgraph Project. In *Proceedings of the 8th Italian Information Retrieval Workshop, Lugano, Switzerland, June 05-07, 2017. (CEUR Workshop Proceedings)*, Vol. 1911. CEUR-WS.org, 121–124. http://ceur-ws.org/Vol-1911/22.pdf
[2] Maurizio Atzori, Simone Balloccu, and Andrea Bellanti. 2018. Unsupervised Singleton Expansion from Free Text. In *12th IEEE International Conference on Semantic Computing, ICSC 2018, Laguna Hills, CA, USA, January 31 - February 2, 2018*. IEEE Computer Society, 180–185. https://doi.org/10.1109/ICSC.2018.00033
[3] José Camacho-Collados, Claudio Delli Bovi, Luis Espinosa Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. 2018. SemEval-2018 Task 9: Hypernym Discovery. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*. Association for Computational Linguistics, 712–724. https://aclanthology.info/papers/S18-1115/s18-1115