The Need of Structured Data: Introducing the *OKgraph* Project
(Extended Abstract)

Maurizio Atzori

Department of Math/CS, University of Cagliari,
Via Ospedale 72, 09124 Cagliari (Italy),
atzori@unica.it
https://git.io/atzori

Abstract. Although many computational problems can be approached using Deep Learning, in this position paper we argue that in the case of Information Retrieval tasks this is not mandatory and even detrimental whenever alternatives exist. Instead of learning (by training) how to solve the full problem, we suggest to split it into two sub-problems: a) producing structured data (specifically knowledge graphs) out of the corpora, and b) providing usable tools (including natural language) to querying such structured data. Motivated by this two-step approach and its need of structured data, we introduce the *Open Knowledge Graph (OKgraph)* project, an initiative recently funded by Regione Autonoma della Sardegna aiming at providing insights on the first part of the problem: a general way of generating knowledge graphs from text corpora, unsupervisedly.

Keywords: word embeddings, knowledge graphs, unsupervised learning, machine understanding

The Need of Structured Data in Information Retrieval

Information Retrieval (IR) is about satisfying the information needs of users. From a very wide perspective, it means providing humans with a simple way to pose questions, for instance through natural language, and then computing the corresponding answer out of some given data, usually unstructured (texts, images, sounds, etc.). Depending on the complexity required to compute the answer, IR can be very challenging, ranging from simple keyword lookup to something close or even harder than passing the *Turing test*. Therefore, from one side we have an unstructured question (e.g., "how many countries in Europe?"), on the other side we have some mostly-unstructured big data to use in order to answer. How to compute answers from the data *is* the problem.

The most innovative way of dealing with virtually any problem, especially those where humans are someway involved such as in IR, is to resort to deep learning. In deep learning, what to do is induced from a vast amount of training data. In the case at hand, supposing the answer is not explicitly available in

the data, the model should explicitly or implicitly learn, among the others, the concept of being a country, associate it to the appropriate entities, separate those in Europe from the others, and then count them.

This research path is disruptive because it generates useful computable functions no programmer is able to code. For instance, a function that given an image can list the objects contained within, or provide a textual description for it. But the greatness of deep learning comes at a cost, actually two drawbacks: on one hand, *a)* the generated function is somewhat a black box, that is, in most of the cases it does not provide any insight or knowledge about the problem; on the other hand *b)* it requires a huge amount of trained data, and for some problems such data simply does not exist yet.

We argue that for the specific research objectives of IR, there is a different research path to approach the problem, that is, exploiting what we already understand well about the problem. By using the same example of counting the countries in Europe, we have that projections, filtering, and aggregates are well-studied (for instance in the Database community) and every programmer can code this functions. In other words, we do not need to learn (using deep learning with its associated drawbacks) what we already know from decades of research in Databases. What we really *need* is to start from what databases need: *structured data*.

Knowledge Graphs

Knowledge Graphs are a machine representation of what Semantic Memories are for humans: general world knowledge that we have accumulated throughout our lives.

For instance, knowing that Rome is (an instance of) a settlement, and it is also the capital of Italy, which is a country. This kind of common sense knowledge is of paramount importance in a number of human tasks such as question answering, disambiguation, understanding, translation, learning, etc. The very same is true for machines, which are able to perform those tasks with high accuracy and recall whenever knowledge graphs about the given context are available.

Many successful crowdsourced projects focused on building *curated* knowledge that can be read by machines: Wikidata, Freebase, DBpedia (machine generated by the humanly-curated infoboxes of Wikipedia), Wordnet. Some other projects focused instead on (automatic) ontology alignment of specific curated knowledge bases, such as BabelNet that produced a multilingual dictionary by aligning parts of Wordnet, Wikidata and other knowledge bases.

All these projects, and their produced knowledge (either "curated-only" or "augmented from curated data") have their merits and very successful applications, but they are also limited by two main weaknesses:

1. they cover only specific (biased) areas of knowledge: very deep and narrow in a few cases (such as LinkedMDB, that knows about movie actors names but not about the number of their children) or wide but shallow (for instance,

none of the previous projects are knowledgeable on, e.g., mobile device features, or recent news, or stock markets);
2. they are pretty much static, that is, once created they do not vary a lot; given the high cost of curating structured data, they tend to avoid time-changing topics (such as news or stock markets). Further, ontologies and automatic procedures (such as alignment) consists of ad-hoc, specific purpose code that is expensive to maintain.

As a consequence, we do not have ontologies and structured data to answer questions such as "what was the average rate on return of italian market on December", or "what's new in last Android release", or "which deputy was the first to sign the divorce law in Italy". These limitations also affect the quality of existing knowledge bases. For instance, according to DBpedia Robbie Williams is a "musical artist" while Lady Gaga is only a "person", therefore not having information on her occupation nor musical genres. Our project SWiPE (Searching WikiPedia by Example) [2, 1], allowing for structured searches in a user-friendly way, has drawn attention to these weaknesses and inconsistencies, clearly showing the need for a more effective way of producing complete and accurate knowledge bases [3].

Therefore, we want to address a currently under-investigated problem: learning machine-readable knowledge graphs from scratch, i.e., without necessarily relying on existing curated knowledge bootstrapping. Like a newborn can learn "from scratch" by listening other's people talk, an empty knowledge graph can be automatically feed with word information coming from natural language texts (such as generalist encyclopedias, news feeds, scientific magazines), opening up a new world of applications which are currently not available or requiring too much efforts to be effective.

The *OKgraph* Project

This project is focused on investigating the fundamental relationship between unstructured data (as natural language text) and structured data (graphs representing knowledge), eventually leading to an autonomous way of inferring the latter from the former. The main outcome of our research will be a computer system, called OKgraph, that learns meaningful graph triples autonomously from scratch, that is, from non-annotated free text such as Wikipedia, WMT11 text data, UMBC WebBase and other available corpora, continuously updating a self-generated knowledge graph (that may resemble DBpedia, Wikidata or Freebase, plus statistics). The approach we are going to follow is based on the exploitation of linear regularities in word vector representations (vectors in a R^N vector space with N indicatively ranging from 300 to 1000) obtained by state-of-the-art word embedding systems. In particular, we are interested in the analogies that such word vectors can represent. For instance, assuming vec(X) is the vector representing the word "X" and vicinity is given by cosine similarity, we have that vec("Rome") - vec("Italy") + vec("France") is closer to vec("Paris") than to any other word vector. Also, vec("Germany") + vec("capital") is close to

vec("Berlin"). This linearity relationship among vectors implies that not only words are semantically correlated through cosine similarity, but information on the specific kind of correlation (concepts relationship such as "being capital of" or "married to") is hidden in the vector representation.

Also inspired by our successful SWiPE System, where structured knowledge graphs (in particular entity-property-value relationships) are merged within text/html data in order to be easily queried through a QBE-like interface [4], we propose to extract entity-property-value data from the word vectors, through techniques to be developed within this project. Vector linearity over different dimensions (in the vector space model) makes our approach very promising.

While current mainstream research efforts are focused on either crowdsourcing (that is humanly-curated, such as Wikidata) or by writing specific parsers for each attribute of interest (DBpedia), using word embeddings to generate knowledge graphs is a novel approach that is expected to be more scalable and more general than existing approaches, with possible disruptive outcomes for the scientific research area.

Learning graphs is a process that will be conduced by using both very large corpora (such as English Wikipedia and books from Gutenberg Project) and medium to small corpora. For the latter, among the others, we want to investigate the use of the Sardinian Wikipedia project[1], currently counting about 5000 entries. We expect many positive outcomes from the project, with possible large impact in the society. Having machine-learned open knowledge graphs allow to populate Wikidata and Wikipedia infoboxes could potentially have a positive impact on less-used languages such as Sardinian or other local languages. More generally, we remark that other than being useful as an interpretable instrument for humans (in contrast with other knowledge representations such as artificial neural networks), knowledge graphs are also beneficial in many areas such as factoid and general question answering, disambiguation, search engines, etc., therefore addressing the need of structured data for IR tasks mentioned in the first part of the paper.

References

1. M. Atzori, S. Gao, G. M. Mazzeo, and C. Zaniolo. Answering end-user questions, queries and searches on wikipedia and its history. *IEEE Data Eng. Bull.*, 39(3):85–96, 2016.
2. M. Atzori and C. Zaniolo. Swipe: searching wikipedia by example. In *WWW 2012*, pages 309–312, 2012. Also featured on New Scientist, see https://www.newscientist.com/article/dn21625-new-search-tool-to-unlock-wikipedia.
3. M. Atzori and C. Zaniolo. Expressivity and accuracy of by-example structured queries on wikipedia. In *24th IEEE WETICE 2015*, pages 239–244, 2015.
4. M. M. Zloof. Query-by-example: The invocation and definition of tables and forms. In *Proceedings of the 1st International Conference on Very Large Data Bases*, VLDB '75, pages 1–24, New York, NY, USA, 1975. ACM.

[1] https://sc.wikipedia.org/