Ontology Learning from the Web

Matúš PIKULIAK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
matus.pikuliak@gmail.com

Data mining from the content of the Web is a field of study that is gaining attention of researches around the world. Vast corpus of data in natural language there contains great bulk of the human knowledge. Texts there are written in virtually every human language and the structure of the Web is enriching them with contextual information. Processing of these texts is a task also known as natural language processing (NLP). Its goal is to understand the meaning of texts written by humans. Complete solution of this task presumably requires construction of so called strong AI – Artificial intelligence comparable to human intelligence with consciousness about itself and outer world.

One of the approaches dealing with natural language processing and knowledge storing is creation of contextual maps, also known as ontologies. Ontologies are knowledge based structures consisting of concepts and relations between them. Process of creating ontologies is also known as ontology learning and as such is a subject of intensive research. In our work we focus on unsupervised, fully automatic ontology learning from unstructured data, which is corpus of texts written in natural language.

Ontology learning is usually done in several steps on several layers with increasing complexity and level of abstraction. Each of these layers has output and several tasks that are essential for achieving it. Traditionally there are four basic layers, but some of them can be further subdivided: terms extraction, concepts acquisition, relations construction, axioms construction [1,3].

In our work we focus on the relations' layer of the ontology learning process. There are several approaches to relations extraction. Good results were achieved using structured data containing information about texts and their subject. We are primarily concerned with statistical analysis. This analysis can be used for finding both taxonomic and non-taxonomic relations between terms and then also concepts. This direction of research is based on so called *distributional hypothesis*, which says that words that tend to occur in similar contexts tend to have similar meaning. Tracking co-occurrence of words in their respective contexts is a basis for so-called word-context matrices, which can be leveraged for finding new relations between words [2].

[*] Supervisor: Marián Šimko, Institute of Informatics and Software Engineering

Spring 2015 PeWe Workshop, April 11, 2015, pp. 79–80.

Approach we are primarily pursuing in our research is based on pair-pattern matrices. These matrices are most suited for measuring semantic similarity between relations. Semantic similarity is for example similarity between relation of Rome an Italy and relation of Paris and France. Both are pair of countries and their respective capital cities. State-of-the-art continuous space language models created from data in natural language can be utilized in working with these relations. Possible applications for pair-pattern matrices in ontology learning are for example relational classifying, relational search, analogical mapping or measuring relational similarity [4].

Continuous space language models represent every word as a vector of various words. Operation with these vectors can be utilized in finding new relations. Following our example with countries and cities we can say that “Rome” minus “Italy” plus “France” is equal to “Paris”. This equation is also depicted in Figure 1. In our scenario we have known there is a relation between Rome and Italy. We have then applied this relation on France and found its capital Paris and therefore we have found new relation between France and Paris.

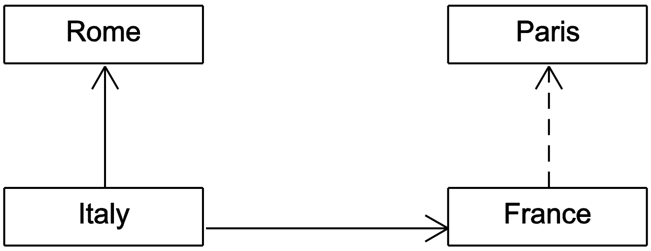

Figure 1. Finding new relations based on latent semantic similarity.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG 1/0646/15.

References

[1] Wong, Wilson; Liu, Wei; Bennamoun, Mohammed. Ontology learning from text: A look back and into the future. ACM Computing Surveys *(CSUR)*, 2012, Volume 44, Issue 4, Article No. 20.

[2] Turney, Peter D., et al. From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research, 2010, Volume 37 Issue 1, pp. 141-188.

[3] Cimiano, Philipp. Ontology learning from text. Springer US, 2006.

[4] Mikolov, Tomas; Yih, Wen-tau; Zweig, Geoffrey. Linguistic Regularities in Continuous Space Word Representations. In: HLT-NAACL. 2013, pp. 746-751.