Hybrid techniques for knowledge-based NLP

Knowledge graphs meet machine learning and all their friends

Jose Manuel Gomez-Perez
Expert System
Madrid, Spain
jmgomez@expertsystem.com

Ronald Denaux
Expert System
Madrid, Spain
rdenaux@expertsystem.com

Daniel Vila
Recogn AI
Madrid, Spain
daniel@recogn.ai

Carlos Badenes
Universidad Politecnica de Madrid
Madrid, Spain
cbadenes@fi.upm.es

ABSTRACT

Many different artificial intelligence techniques can be used to explore and exploit large document corpora that are available inside organizations and on the Web. While natural language is symbolic in nature and the first approaches in the field were based on symbolic and rule-based methods, like ontologies, semantic networks and knowledge bases, most widely used methods are currently based on statistical approaches, including linear methods, such as support vectors machines or probabilistic topic models, and non-linear ones such as neural networks. Each of these two main schools of thought in natural language processing, knowledge-based and statistical, have their limitations and strengths and there is an increasing trend that seeks to combine them in complementary ways to get the best of both worlds. This tutorial will cover the foundations and modern practical applications of knowledge-based and statistical methods, techniques and models and their combination for exploiting large document corpora. The tutorial will first focus on the foundations of many of the techniques that can be used to this purpose, including knowledge graphs, word embeddings, neural network methods, and probabilistic topic models, and will then show how these techniques are being effectively combined in practical applications, including commercial projects where the instructors currently participate.

KEYWORDS

Knowledge graphs, Hybrid natural language processing, embeddings, vecisgrafo, topic models

1 MOTIVATION

For several decades, semantic systems were predominantly developed around knowledge graphs at different degrees of expressivity. Through the explicit representation of knowledge in well-formed, logically sound ways, knowledge graphs provide knowledge-based text analytics with rich, expressive and actionable descriptions of the domain of interest and support logical explanations of reasoning outcomes. On the downside, knowledge graphs can be costly to produce since they require a considerable amount of human effort to manually encode knowledge in the required formats. Additionally, such knowledge representations can sometimes be excessively rigid and brittle in the face of different natural language processing applications, like e.g. question answering.

In parallel, the last decade has witnessed a shift towards statistical methods to text understanding due to the increasing availability of raw data and cheaper computing power. Such methods have proved to be powerful and convenient in many linguistic tasks. Particularly, recent results in the field of distributional semantics have shown promising ways to learn language models from text, encoding the meaning of each word in the corpus as a vector in dense, low-dimensional spaces. Among their applications, word embeddings have proved to be useful in term similarity, analogy and relatedness, as well as many downstream tasks in natural language processing.

Aimed towards Semantic Web researchers and practitioners, this tutorial elaborates on the idea introduced in [1] and shows how it is possible to bridge the gap between knowledge-based and statistical approaches to further knowledge-based natural language processing. Following a practical and hands-on approach, the tutorial tries to address a number of fundamental questions to achieve this goal, including: How can Machine Learning techniques be used to complement the knowledge already captured explicitly in knowledge graphs, extending and curating them in cost-efficient and practical ways, what are the main building blocks and techniques enabling such hybrid approach to natural language processing, how can structured and statistical knowledge representations be seamlessly integrated, how can the quality of the resulting hybrid representations be inspected and evaluated, and how can this improve the overall quality and coverage of our knowledge graphs.

2 DESCRIPTION OF THE TUTORIAL

This half-day tutorial provides plenty of practical content, real-life examples and applications, and exercises. We offer an interactive session where both instructors and participants can engage in rich discussions on the topic. The agenda addresses the following points.

- Probabilistic topic models and topic-based semantic similarity.
- Creating a language model through word embeddings.
- Extending word embeddings with structured knowledge.
- Creating knowledge graph embeddings.
- Building a vecsigrafo - bringing knowledge from text into knowledge graphs.

K-CAP2017 Workshops and Tutorials Proceedings, 2017

- Evaluating vecsigrafos beyond visual inspection and intrinsic methods.
- Applications in cross-lingual natural language processing.
- Putting it all together in a real-life system.
- Beyond text understanding: Cross-modal extensions.

3 MATERIALS

The tutorial follows a highly practical approach. The teaching fundamentally consist of Jupyter notebooks that participants can install locally through Docker images with all the necessary software to run the examples and exercises in their own machines. The materials of the K-CAP 2017 tutorial can be found in GitLab[1]

4 AUDIENCE

This tutorial seeks to be of special value for members of the Semantic Web community although it is also useful for related communities, e.g. Machine Learning and Computational Linguistics. We welcome researchers and practitioners both from industry and academia, as well as other participants with an interest in hybrid approaches to knowledge-based natural language processing.

5 PRESENTERS

The tutorial is offered by the following members instructors.

Jose Manuel Gomez-Perez works in the intersection of several areas of Artificial Intelligence, including Natural Language Processing, Knowledge Discovery, Representation and Reasoning. His long-term vision is to enable machines to understand text in a way similar to how humans read, bridging the gap between both through semantically rich knowledge representations and user interfaces. At Expert System, Jose Manuel leads the Research Lab in Madrid where he focuses on the combination of structured knowledge graphs and probabilistic methods. Before Expert System, he worked at iSOCO, one of the first European companies to deliver semantic and natural language processing solutions on the Web. He consults for companies like Coca-Cola or ING. Also active as an entrepreneur, he co-founded a startup and advised another. An ACM member and Marie Curie fellow, Jose Manuel holds a Ph.D. in Computer Science and AI from UPM and regularly publishes in top scientific conferences and journals. His views on AI and applications have appeared in magazines like Nature and Scientific American. In 2015, he was the program chair of the International Conference on Knowledge Capture (K-CAP).

Ronald Denaux is a senior researcher at Expert System. Ronald obtained his MSc in Computer Science from the Technical University Eindhoven, The Netherlands. After a couple of years working in industry as a software developer for a large IT company in The Netherlands, Ronald decided to go back to academia. He obtained a PhD, again in Computer Science, from the University of Leeds, UK. Ronald's research interests have revolved around making semantic web technologies more usable for end users, which has required research into (and resulted in various research publications in) the areas of Ontology Authoring and Reasoning, Natural Language Interfaces, Dialogue Systems, Intelligent User Interfaces and User Modelling. Besides research, Ronald also participates in knowledge transfer and product development.

[1] https://gitlab.com/rdenaux/kcap17-tutorial

Daniel Vila is co-founder of recogn.ai, a Madrid-based startup and spin-off from UPM, building next generation solutions for text analytics and content management using the AI methods. Daniel holds a PhD in Artificial Intelligence by Universidad PolitÃĺcnica de Madrid (2016), where he worked at the Ontology Engineering Group and developed the solution supporting a large knowledge graph combining NLP and semantic technologies: the datos.bne.es data service from the National Library of Spain.

Carlos Badenes: After more than 8 years working on the M2M world, Carlos began researching about text mining within the context of the Semantic Web. Since then, he has moved more deeply into the study of topic modeling techniques to analyze large collections of documents, incorporating semantic resources and working on multilingual domains. He currently works as an associate researcher at the Ontology Engineering Group doing a PhD at UPM.

Oscar Corcho: Oscar Corcho is Full Professor at Departamento de Inteligencia Artificial, UPM, and belongs to the Ontology Engineering Group. His research is focused on Semantic e-Science and Real World Internet, although he also works in more general areas of Semantic Web and Ontological Engineering. He has participated in numerous EU and Spanish R&D projects as well as privately-funded projects like ICPS (International Classification of Patient Safety), funded by the World Health Organisation, and HALO, funded by Vulcan Inc. Previously, he worked as a Marie Curie research fellow at the University of Manchester, and was a research manager at iSOCO. He holds a PhD in Computer Science and AI from UPM. He was awarded the Third National Award by the Spanish Ministry of Education in 2001. He has published several books, from which âĂIJOntological Engineering" can be highlighted as it is being used as a reference book in a good number of university lectures worldwide, and more than 100 papers in journals, conferences and workshops. He usually participates in the organization or in the program committees of relevant international conferences and workshops.

ACKNOWLEDGMENTS

Partially funded by the EU H2020 project DANTE (700367) and the national Spanish project GRESLADIX (20160805).

REFERENCES

[1] Ronald Denaux and Jose Manuel Gomez-Perez. 2017. Towards a vecsigrafo: Portable semantics in knowledge-based text analytics.. In *Proceedings of the 2017 workshop on Hybrid Statistical Semantic Understanding and Emerging Semantic (HSSUES '17)*. Held in conjunction with the 16th Intl. Semantic Web Conference, CEUR Workshop Proceedings.