A Practical Guide to Hybrid Natural Language Processing

Jose Manuel Gomez-Perez • Ronald Denaux •
Andres Garcia-Silva

A Practical Guide to Hybrid Natural Language Processing

Combining Neural Models and Knowledge Graphs for NLP

Springer

Jose Manuel Gomez-Perez
Expert System
Madrid, Spain

Ronald Denaux
Expert System
Madrid, Spain

Andres Garcia-Silva
Expert System
Madrid, Spain

ISBN 978-3-030-44829-5 ISBN 978-3-030-44830-1 (eBook)
https://doi.org/10.1007/978-3-030-44830-1

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To family and friends and all who made this book possible in one way or another.

Foreword

"Don't Read This Book. Use It!" by Ken Barker

For as long as Knowledge-based Language Understanding and Statistical Natural Language Processing have coexisted, we have fought over them. "Statistical NLP is superficial! It is not real understanding. It will never offer more than parlor tricks learned from common patterns in text." But "Knowledge heavy approaches are brittle! They rely on arbitrary, expensive, hand-coded rules that are not flexible enough to deal with the incredible variety found in real-world text. Besides, you can do anything with BERT."

Both of these positions are caricatures, and both contain some truth. But they are barriers to exploring all of the available tools that may be needed *just to get the job done.* By focusing on the limitations of symbolic or data-driven approaches, we risk losing out on the unique advantages each can offer, as well as the even greater power of combining them.

A Practical Guide to Hybrid Natural Language Processing does not belong in either camp. It devotes no space to ancient wars and is not intended for an audience interested in rehashing old arguments. It *is* intended for those coming from a background of Symbolic AI; those who have been following the impressive successes of Statistical and now Neural NLP, but have not yet taken the plunge into embeddings, language models, transformers, and Muppets. The book is also for those statistical NLP practitioners who have struggled against the insatiable appetite of modern techniques for data; those who suspect that adding knowledge might enable learning from fewer examples, if only there were an effective way to incorporate it. More practically, this book is ideal for those who want to build something useful that requires getting at the meaning locked in text and language and who don't want to have to get a Ph.D. in Logic or spend Googles of dollars on GPU time.

The authors have taken care to include thorough treatments of all of the basic components for building hybrid NLP systems that combine the power of knowledge graphs with modern, neural techniques. This is not to say that the book is an

encyclopedia of techniques, either data-oriented or symbolic. It is also not a textbook, dragging you through a fixed path of education. The book is divided into three parts: knowledge-based and neural building blocks; hybrid architectures combining both; and real applications. Within the parts, topics are conveniently standalone, allowing quick, easy access to needed information. But the two most valuable properties of the book are that it is practical and that it is up to date. It demonstrates exactly how to create and use contextual representations. It has clear treatments of sense embeddings and knowledge graph embeddings. It explains language models and transformer architectures that use them. It also shows how to evaluate the performance of systems using these. Most usefully, the book takes you from theory to code as painlessly as possible through experiments and exercises on real NLP tasks in real domains with real corpora. It is packed with working code and step-by-step explanations. And it uses Jupyter notebooks with pandas that you can get from GitHub!

My advice is: Don't read this book. Use it! Work through its experiments and exercises. Step through the notebooks and see what happens. Then steal the code and build the NLP system you need.

IBM Research
Yorktown Heights, NY, USA
February 2020

Ken Barker

"Most of the Knowledge in the World Is Encoded Not in Knowledge Graphs but in Natural Language" by Denny Vrandecic

In 2005, Wikipedia was still a young website, and most of the public just started to become aware of it. The Wikipedia community sent out calls for the very first global meet-up of contributors, named Wikimania. Markus Krotzsch and I were both early contributors and have just started as Ph.D. students working on the Semantic Web. We wanted to go to Wikimania and meet those people we knew only online.

We sat down and thought about what kind of idea to submit to Wikimania. The most obvious one was to combine Wikipedia with Semantic Web technologies. But what would that mean?

Wikipedia's power lies in the ease of editing, and in the idea that anyone can contribute to it. It lies in its community, and in the rules and processes the community had set up. The power of the Semantic Web was to publish machine-readable data on the Web and to allow agents to combine the data from many different sources. Our idea was to enable Wikipedia communities to create content that is more machine-readable and can participate in the Semantic Web.

Our talk was set up for the first session on the first day, and we used the talk to start a conversation that would continue for years. The talk led to the creation of Semantic MediaWiki, an extension to the MediaWiki software powering Wikipedia

Fig. 1 Map showing the 100 largest cities with a female mayor. Data from Wikidata, Map from Open StreetMaps. Link: https://w.wiki/GbC

and other wikis, and that has found use in numerous places, such as NASA, the Museum of Modern Art, the US Intelligence community, General Electric, and many more. The talk also eventually led to the creation of Wikidata, but it took many years to get there.

In the talk, we proposed a system that would allow us to answer questions using Wikipedia's content. Our example question was: what are the world's largest cities with a female mayor?

Wikipedia, at that point, already had all the relevant data, but it was spread out through many articles. One could, given enough time, comb through the articles of all the largest cities, check who the mayor is, and start keeping a spreadsheet of potential answers, and finally clean it up and produce the end result.

Today, with the availability of Wikidata, we can get answers to that question (see Fig. 1) and many others within seconds. Wikidata is a large knowledge base that anyone can edit and that has, as of early 2020, collected nearly a billion statements about more than 75 million topics of interest.

But, although we finally have a system that allows all of us to ask these questions and get beautiful visualizations as answers, there is still a high barrier towards allowing a wide range of people to actually benefit from this data. It requires writing queries in the query language SPARQL, a rare skill set. Can we do better?

Most of the knowledge in the world is encoded not in knowledge graphs but in natural language. Natural language is also the most powerful user interface we know.

The 2010s saw neural models created through deep learning applied to tasks in natural language processing become hugely popular. Whether generation, summarization, question answering, or translation—undoubtedly the poster child of this

development—neural models have turned from an interesting research idea to the foundation of a multi-billion dollar market.

At the same time, large knowledge graphs have become widely available and blossomed. Public projects such as DBpedia, Freebase, and Wikidata are widely used, while a growing number of companies have built their internal knowledge graphs. Foremost Google, whose Knowledge Graph popularized the name, but today many large companies across diverse industries have internal knowledge graphs used in diverse ways.

Both technologies, machine learning and knowledge graphs, have evolved much over the last few years. The opportunities coming from their integration are underexplored and very promising. It is natural language processing that can help ensure that the content of Wikidata is indeed supported by the text of Wikipedia and other referenced sources, and it is natural language processing that can enable a much larger population to access the rich knowledge within a knowledge base like Wikidata.

This is where this book comes into play. Jose, Ronald, and Andres are experts in their fields, with decades of experience shared between them. But, in order to write this book, they had to resolve the hard problems of aligning terminology and how to present the ideas from both sides in a unified framework that is accessible to all kinds of readers. They are generously linked to resources on the Web that let you try out the techniques described in the book. The content of the book has been tested in tutorials and refined over many months.

Natural language is, without any doubt, one of the most important user interface modalities of the future. Whereas we see a growing number of devices that aim to converse with us in natural language, this is only the very beginning of a massive revolution. Natural language interfaces provide extremely powerful and intuitive methods to access the ever-growing capabilities of computer systems. Already today, the services computers could offer are far behind the services that are, in fact, being offered to the user. There is so much more users could do, but we do not know how to build the user interface to these powerful capabilities.

Knowledge graphs provide the most interpretable and efficient way to store heterogeneous knowledge we are aware of today. Knowledge graphs allow for human interpretation and editing capabilities. In a knowledge graph-based system, you can dive in and make a change, and be sure that this change will propagate to the users. In many industries, such as finance or medicine, this is not only a nice to have, but absolutely crucial. And in other industries, such requirements are starting to become increasingly important. Knowledge graphs offer an ease of maintenance and introspection that outmatches many alternatives.

Natural language processing is unlocking knowledge expressed in natural language, which can then be used to update and be checked for consistency regarding a knowledge graph, which in turn can be made available through natural language for

querying and answer generation. The two technologies support and enhance each other. This book shows practical ways to combine them and unlock new capabilities to engage and empower users.

Enjoy the book, and use your newly acquired knowledge wisely!

Google Knowledge Graph
San Francisco, CA, USA
January 2020

Denny Vrandecic

Preface

Both neural and knowledge-based approaches to natural language processing have strong and weak points. Neural methods are extremely powerful and consistently claim the top positions of current NLP leaderboards. However, they are also sensitive to challenges like the amount and quality of training data or linking the models to how humans use the language and their understanding of the world. On the other hand, although not entirely free from such challenges, NLP systems based on structured knowledge representations tend to be better suited to address some of them. However, they may require considerable knowledge engineering work in order to continuously curate such structured representations.

The main premise of this book is that data-driven and knowledge-based approaches can complement each other nicely to boost strengths and alleviate weaknesses. Although many advocate for the combined application of both paradigms in NLP and many other areas of Artificial Intelligence, the truth is that until now such combination has been unusual, due to reasons that may include a possible lack of principled approaches and guidelines to accomplish such goal and a shortage of compelling success stories.

On the other hand, AI research, especially in the areas of NLP and knowledge graphs, has reached a level of maturity that permeates all sectors, causing profound societal and business changes. Therefore, this book focuses particularly on the practical side of the topics discussed herein and aims to provide the interested reader with the necessary means to acquire a hands-on understanding about how to combine neural and knowledge-based approaches to NLP, bridging the gap between them.

In general, this book seeks to be of value for anyone interested in the interplay between neural and knowledge-based approaches to NLP. Readers with a background on structured knowledge representations from the Semantic Web, knowledge acquisition, representation, and reasoning communities, and in general those whose main approach to AI is fundamentally based on logic can find in this book a useful and practical guide. Likewise, we expect it to be similarly useful for readers from communities whose main background is in the areas of machine and deep learning,

who may be looking for ways to leverage structured knowledge bases to optimize results along the NLP downstream.

Readers from industry and academia in the above-mentioned communities will thus find in this book a practical resource to hybrid NLP. Throughout the book, we show how to leverage complementary representations stemming from the analysis of unstructured text corpora as well as the entities and relations described explicitly in a knowledge graph, integrate such representations, and use the resulting features to effectively solve different NLP tasks in different domains. In these pages, the reader will have access to actual executable code with examples, exercises, and real-world applications in key domains like disinformation analysis and machine reading comprehension of scientific literature.

In writing this book, we did not seek to provide an exhaustive account of current NLP approaches, techniques, and toolkits, either knowledge-based, neural, or based on other forms of machine learning. We consider this is sufficiently well covered in the literature. Instead, we chose to focus on the main building blocks that the reader actually needs to be aware of in order to assimilate and apply the main ideas of this book. Indeed, all the chapters are self-contained and the average reader should not encounter major difficulties in their comprehension. As a result, you have in your hands a compact yet insightful handbook focused on the main challenge of reconciling knowledge-based and neural approaches to NLP. We hope you will enjoy it.

Madrid, Spain
January 2020

Jose Manuel Gomez-Perez
Ronald Denaux
Andres Garcia-Silva

Purpose of the Book

This book provides readers with a principled yet practical guide to hybrid approaches to natural language processing involving a combination of neural methods and knowledge graphs. The book addresses a number of questions related to hybrid NLP systems, including:

- How can neural methods extend previously captured knowledge explicitly represented as knowledge graphs in cost-efficient and practical ways and vice versa?
- What are the main building blocks and techniques enabling a hybrid approach to NLP that combines neural and knowledge-based approaches?
- How can neural and structured, knowledge-based representations be seamlessly integrated?
- Can this hybrid approach result in better knowledge graphs and neural representations?

- How can the quality of the resulting hybrid representations be inspected and evaluated?
- What is the impact on the performance of NLP tasks, the processing of other data modalities, like images or diagrams, and their interplay?

To this purpose, the book first introduces the main building blocks and then describes how they can be intertwined, supporting the effective implementation of real-life NLP applications. To illustrate the ideas described in the book, we include a comprehensive set of experiments and exercises involving different algorithms over a selection of domains and corpora in several NLP tasks.

Overview of the Chapters in This Book

We have structured the book in the chapters introduced next.

Chapter 1: Introduction motivates the book and its overall purpose in the current context of the NLP discipline.

Chapter 2: Word, Sense, and Graph Embeddings introduces word, sense/concept, and knowledge graph embeddings as some of the main building blocks towards producing hybrid NLP systems. Different approaches are considered, varying from those learning plain word embeddings, to those learning sense and concept embeddings from corpora and semantic networks, and those which do not use corpora at all, but instead attempt to learn concept embeddings directly from a knowledge graph.

Chapter 3: Understanding Word Embeddings and Language Models focuses on word embeddings and delves in the analysis of the information contained in them, depending on the method and corpora used. Beyond pre-trained static embeddings, the emphasis is placed on neural language models and contextual embeddings.

Chapter 4: Capturing Meaning from Text as Word Embeddings guides the reader through an executable notebook, which focuses on a specific word embedding algorithm like Swivel [164] and its implementation to illustrate how word embeddings can be easily generated from text corpora.

Chapter 5: Capturing Knowledge Graph Embeddings. In a way analogous to the previous chapter, this chapter takes an existing knowledge graph like WordNet to produce graph embeddings using a specific knowledge graph algorithm like HolE. An executable notebook is also provided.

Chapter 6: Building Hybrid Knowledge Representations from Text Corpora and Knowledge Graphs presents **Vecsigrafo** [39], an approach to jointly learn word and concept embeddings from text corpora using knowledge graphs. As opposed to the methods described in the previous chapter, Vecsigrafo not only learns from the knowledge graph but also from the training corpus, with several advantages, as we will see, which are illustrated in an accompanying notebook. In the second half of the chapter, we take a step further and show how to apply transformers and neural language models to generate an analogous representation

of Vecsigrafo, called **Transigrafo**. This part of the chapter is also illustrated using a notebook.

Chapter 7: Quality Evaluation discusses several evaluation methods that provide an insight on the quality of the hybrid representations learnt by Vecsigrafo. To this purpose, we will use a notebook that illustrates the different techniques entailed. In this chapter, we also study how such representations compare against lexical and semantic embeddings produced by other algorithms.

Chapter 8: Capturing Lexical, Grammatical, and Semantic Information with Vecsigrafo. Building hybrid systems that leverage both text corpora and a knowledge graph needs to generate embeddings for the items represented in the graph, such as concepts, which are linked to the words and expressions in the corpus singled out through some tokenization strategy. In this chapter and associated notebook, we investigate the impact of different tokenization strategies and how these may impact on the resulting lexical, grammatical, and semantic embeddings in Vecsigrafo.

Chapter 9: Aligning Embedding Spaces and Applications for Knowledge Graphs presents several approaches to align the vector spaces learned from different sources, possibly in different languages. We discuss various applications such as multi-linguality and multi-modality, which we also illustrate in an accompanying notebook. The techniques for vector space alignment are particularly relevant in hybrid settings, as they can provide a basis for knowledge graph interlinking and cross-lingual applications.

Chapter 10: A Hybrid Approach to Fake News and Disinformation Analysis. In this chapter and corresponding notebooks, we start looking at how we can apply hybrid representations in the context of specific NLP tasks and how this improves the performance of such tasks. In particular, we will see how to use and adapt deep learning architectures to take into account hybrid knowledge sources to classify documents which in this case may contain misinformation.

Chapter 11: Jointly Learning Text and Visual Information in the Scientific Domain. In this chapter and its notebook, we motivate the application of hybrid techniques to NLP in the scientific domain. This chapter will guide the reader to implement state-of-the-art techniques that relate both text and visual information, enrich the resulting features with pre-trained knowledge graph embeddings, and use the resulting features in a series of transfer learning tasks, ranging from figure and caption classification to multiple-choice question answering over text and diagram of 6th grade science questions.

Chapter 12: Looking Into the Future of Natural Language Processing provides final thoughts and guidelines on the matter of this book. It also advances some of the future developments in hybrid natural language processing in order to help professionals and researchers configure a path of ongoing training, promising research fields, and areas of industrial application. This chapter includes feedback from experts in areas related to this book, who were asked about their particular vision, foreseeable barriers, and next steps.

Materials

All the examples and exercises proposed in the book are available as executable Jupyter notebooks in our GitHub repository.[1] All the notebooks are ready to be run on Google Colaboratory or, if the reader so prefers, in a local environment. The book also leverages experience and feedback acquired through our tutorial on *Hybrid Techniques for Knowledge-based NLP*,[2] initiated at K-CAP'17[3] and continued in ISWC'18[4] and K-CAP'19.[5] The current version of the tutorial is available online and the reader is encouraged to use it in combination with the book to consolidate the knowledge acquired in the different chapters with executable examples, exercises, and real-world applications.

Relation to Other Books in the Area

The field addressed by this book is tremendously dynamic. Much of the relevant bibliography in critical and related areas like neural language models has erupted in the last months, configuring a thriving field which is taking shape as we write these lines. New and groundbreaking contributions are therefore expected to appear during the preparation of this book that will be studied and incorporated and may even motivate future editions. For this reason, resources like the above-mentioned tutorial on *Hybrid Techniques for Knowledge-based NLP* and others like Graham Neubig et al.'s *Concepts in Neural Networks for NLP*[6] are of particular importance.

This book does not seek to provide an exhaustive survey on previous work in NLP. Although we provide the necessary pointers to the relevant bibliography in each of the areas we discuss, we have purposefully kept it succinct and focused. Related books that will provide the reader with a rich background in relevant areas for this book include the following.

Manning and Schutze's *Foundations of Statistical Natural Language Processing* [114] and Jurafsky and Martin's *Speech and Language Processing* [88] provide excellent coverage for statistic approaches to natural language processing and their applications, as well as introduce how (semi)structured knowledge representations and resources like WordNet and FrameNet [12] can play a role in the NLP pipeline.

More recently, a number of books have covered the field with special emphasis on neural approaches. Eisenstein's *Introduction to Natural Language Processing* [51]

[1] https://github.com/hybridnlp/tutorial.

[2] http://hybridnlp.expertsystemlab.com/tutorial.

[3] 9th International Conference on Knowledge Capture (https://www.k-cap2017.org).

[4] 17th International Semantic Web Conference (http://iswc2018.semanticweb.org).

[5] 10th International Conference on Knowledge Capture (http://www.k-cap.org/2019).

[6] https://github.com/neulab/nn4nlp-concepts.

offers a comprehensive and up-to-date survey of the computational methods necessary to understand, generate, and manipulate human language, ranging from classical representations and algorithms to contemporary deep learning approaches. Also of particular interest to acquire a deep and practical understanding of the area is Goldberg's *Neural Network Methods in Natural Language Processing* [67].

We also look at books that pay special attention to the knowledge-based side of the NLP spectrum like Cimiano et al.'s *Ontology-Based Interpretation of Natural Language* [33] and Barrière's *Natural Language Understanding in a Semantic Web Context* [17]. A good overview on the application of distributed representations in knowledge graphs is provided by Nickel et al. in [129].

Relevant books on knowledge graphs include Pan et al.'s *Exploiting Linked Data and Knowledge Graphs in Large Organisations* [135], which addresses the topic of exploiting enterprise linked data with a particular focus on knowledge graph construction and accessibility. Kejriwal's *Domain-Specific Knowledge Graph Construction* [91] also focuses on the actual creation of knowledge graphs. Finally, *Ontological Engineering* [68] by Asuncion Gomez-Perez et al. provides key principles and guidelines for the tasks involved in knowledge engineering.

Acknowledgements

We gratefully acknowledge the European Language Grid-825627 and Co-inform-770302 EU Horizon 2020 projects, as well as previous grants, including DANTE-700367, TRIVALENT-740934, and GRESLADIX-IDI-20160805, whose research challenges related to different areas of natural language processing encouraged us to find solutions based on the combination of knowledge graphs and neural models. We are especially thankful to Flavio Merenda, Cristian Berrio, and Raul Ortega for their technical contributions to this book.

Contents