Word Vector Space Specialisation

EACL 2017 Tutorial

Ivan Vulić[1], Nikola Mrkšić[2], Mohammad Taher Pilehvar[1]
[1] Language Technology Lab, DTAL, University of Cambridge
[2] Dialogue Systems Group, Department of Engineering, University of Cambridge
iv250@cam.ac.uk nm480@cam.ac.uk mp792@cam.ac.uk

1 Motivation and Objectives

Specialising vector spaces to maximise their content with respect to one key property of vector space models (e.g. semantic similarity vs. relatedness or lexical entailment) while mitigating others has become an active and attractive research topic in representation learning. Such specialised vector spaces support different classes of NLP problems. Proposed approaches fall into two broad categories: a) Unsupervised methods which learn from raw textual corpora in more sophisticated ways (e.g. using context selection, extracting co-occurrence information from word patterns, attending over contexts); and b) Knowledge-base driven approaches which exploit available resources to encode external information into distributional vector spaces, injecting knowledge from semantic lexicons (e.g., WordNet, FrameNet, PPDB). In this tutorial, we will introduce researchers to state-of-the-art methods for constructing vector spaces *specialised* for a broad range of downstream NLP applications. We will deliver a detailed survey of the proposed methods and discuss best practices for intrinsic and application-oriented evaluation of such vector spaces.

Throughout the tutorial, we will provide running examples reaching beyond English as the only (and probably the easiest) use-case language, in order to demonstrate the applicability and modelling challenges of current representation learning architectures in other languages.

2 Tutorial Overview

2.1 Part I: Learning from Context

First, we will provide a brief overview of the most popular representation learning architectures in NLP which serve as a typical initialisation point for semantic specialisation models. Following that, we will discuss generalisations of these models that can learn from arbitrary contexts reaching beyond the bag-of-words assumption. We will show how to train unsupervised representation models using more informed positional, dependency-based, attention-based, and multilingual contexts, and demonstrate how the choice of the context type affects the properties of induced semantic spaces. We will also describe class-specialised models supported by template-based approaches (e.g., symmetric patterns).

2.2 Part II: Learning from Knowledge Bases

We will present the methods which go beyond learning from raw textual corpora, now relying on human- or automatically- constructed knowledge bases to enrich the embedded content of existing word vector collections. We will place an emphasis on understanding vector space specialisation as both a natural extension to the pre-training steps ubiquitous across deep learning pipelines, and as an elegant mathematical mechanism for *encoding external knowledge* into statistical NLP frameworks. We will describe two distinct approaches to injecting external information: a) the *heavy-weight* ones which jointly learn from textual corpora and embed external constraints; and b) the *post-processing* approaches which inject semantic constraints while trying to preserve beneficial content from the initial vector spaces.

2.3 Part III: Types of Specialisation

We will then discuss other specialisation modalities (i.e. lexical entailment, relatedness, antonymy), and introduce the construction of cross-lingual vector spaces as (universal) semantic specialisation. We will analyse how different target specialisations affect the output of specialised semantic spaces: these insights will demonstrate that the specialisation process may be observed as an extra refinement of initial pre-trained vectors, making them better suited for downstream NLP tasks.

2.4 Part IV: Evaluation and Applications

We will give an overview of existing intrinsic evaluation metrics used for assessing the quality of word vector spaces (SimLex-999, WordSim-353, SimVerb, HyperLex, WS-353, etc.), including a brief survey of state-of-the-art models for each of these datasets. We will analyse several downstream applications which benefit from the use of specialised vector spaces, including Question Answering, Dialogue State Tracking, Natural Language Inference, and Machine Translation. By placing special emphasis on the correlation between downstream performance and widely used intrinsic evaluation metrics, we hope to encourage better practices and use of meaningful evaluation metrics for future work on learning word vector space representations in general and specialised vector spaces in particular.

2.5 Final Remarks

We will conclude by listing publicly available software packages and implementations, available training datasets, and evaluation protocols, as well as providing directions for future work and remaining research challenges in this area.

3 Structure

- **Part I:** Learning from Context (*40 minutes*)
- **Part II:** Learning from Knowledge Bases (*50 minutes*)
- **Coffee Break** (*30 minutes*)
- **Part III:** Types of Specialisation (*30 minutes*)
- **Part IV:** Evaluation and Applications (*45 minutes*)
- **Final Remarks** (*15 minutes*)

4 About the Speakers

Ivan Vulić (https://sites.google.com/site/ivanvulic/) is a research associate at the University of Cambridge. He received his PhD *summa cum laude* at KU Leuven in 2014. Ivan is interested in representation learning, distributional and multi-modal semantics in monolingual and multilingual contexts, and transfer learning for enabling cross-lingual NLP applications. His work has been published in top-tier *ACL and *IR conferences. He gave a tutorial on topic models at ECIR 2013 and WSDM 2014, and co-organised a Vision & Language workshop at EMNLP 2015.

Nikola Mrkšić (mi.eng.cam.ac.uk/~nm480) is a final-year PhD student at the University of Cambridge, supervised by Steve Young. Nikolas research is focused on belief tracking in human-machine dialogue, specifically in moving towards building open-domain language understanding models that are fully data-driven. He is also interested in deep learning, semantics, Bayesian nonparametrics, unsupervised and semi-supervised learning. He has authored several papers in top NLP conferences.

Mohammad Taher Pilehvar (http://people.ds.cam.ac.uk/mp792/) is a research associate at the University of Cambridge. Tahers research lies in lexical semantics, mainly focusing on semantic representation, semantic similarity, and WSD. He has co-organised three SemEval tasks and has authored multiple conference and journal papers on semantic representation in top tier venues (including an ACL 2013 best paper nomination). Taher has co-instructed two well attended tutorials at EMNLP 2015 (on semantic similarity) and ACL 2016 (on sense representations). He is also the co-organiser of an EACL 2017 workshop on fine-grained representations.