Embeddings in Natural Language Processing

Jose Camacho-Collados[1] and **Mohammad Taher Pilehvar**[2,3]
[1]School of Computer Science and Informatics, Cardiff University, UK
[2]Tehran Institute for Advanced Studies (TeIAS), Tehran, Iran
[3]DTAL, University of Cambridge, UK
camachocolladosj@cardiff.ac.uk, mp792@cam.ac.uk

Abstract

Embeddings have been one of the most important topics of interest in Natural Language Processing (NLP) for the past decade. Representing knowledge through a low-dimensional vector which is easily integrable in modern machine learning models has played a central role in the development of the field. Embedding techniques initially focused on words but the attention soon started to shift to other forms. This tutorial will provide a high-level synthesis of the main embedding techniques in NLP, in the broad sense. We will start by conventional word embeddings (e.g., Word2Vec and GloVe) and then move to other types of embeddings, such as sense-specific and graph alternatives. We will finalize with an overview of the trending contextualized representations (e.g., ELMo and BERT) and explain their potential and impact in NLP.

1 Description

In this tutorial we will start by providing a historical overview on word-level vector space models, and word embeddings in particular. Word embeddings (e.g. Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) or FastText (Bojanowski et al., 2017)) have proven to be powerful keepers of prior knowledge to be integrated into downstream Natural Language Processing (NLP) applications.

However, despite their flexibility and success in capturing semantic properties of words, the effectiveness of word embeddings are generally hampered by an important limitation, known as the meaning conflation deficiency: the inability to discriminate among different meanings of a word. A word can have one meaning (monosemous) or multiple meanings (ambiguous). For instance, the noun mouse can refer to two different meanings depending on the context: an animal or a computer device. Hence, mouse is said to be ambiguous. In fact, according to the Principle of Economical Versatility of Words (Zipf, 1949), frequent words tend to have more senses. Moreover, this meaning conflation can have additional negative impacts on accurate semantic modeling, e.g., semantically unrelated words that are similar to different senses of a word are pulled towards each other in the semantic space (Neelakantan et al., 2014; Pilehvar and Collier, 2016). In our example, the two semantically-unrelated words rat and screen are pulled towards each other in the semantic space for their similarities to two different senses of mouse (see Figure 1). This, in turn, contributes to the violation of the triangle inequality in euclidean spaces (Tversky and Gati, 1982; Neelakantan et al., 2014).

Accurately capturing the meaning of words (both ambiguous and unambiguous) plays a crucial role in the language understanding of NLP systems. In order to deal with the meaning conflation deficiency, this tutorial covers approaches have attempted to model individual word senses (Reisinger and Mooney, 2010; Huang et al., 2012; Neelakantan et al., 2014; Rothe and Schütze, 2015; Li and Jurafsky, 2015; Pilehvar and Collier, 2016; Mancini et al., 2017). Sense representation techniques, however, suffer from limitations which hinders their effective application in downstream NLP tasks: they either need vast amounts of training data to obtain reliable representations or require an additional sense disambiguation on the input text to make them integrable into NLP systems. This data is highly expensive to obtain in practice, which causes the so-called knowledge-acquisition bottleneck (Gale et al., 1992).

As a practical way to deal with the knowledge-acquisition bottleneck, an emerging branch of research has focused on directly integrating unsupervised embeddings into downstream models. Instead of learning a fixed number of senses per word, contextualized word embeddings learn “senses” dynamically,

Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts, pages 10–15
Barcelona, Spain (Online), December 12th, 2020.

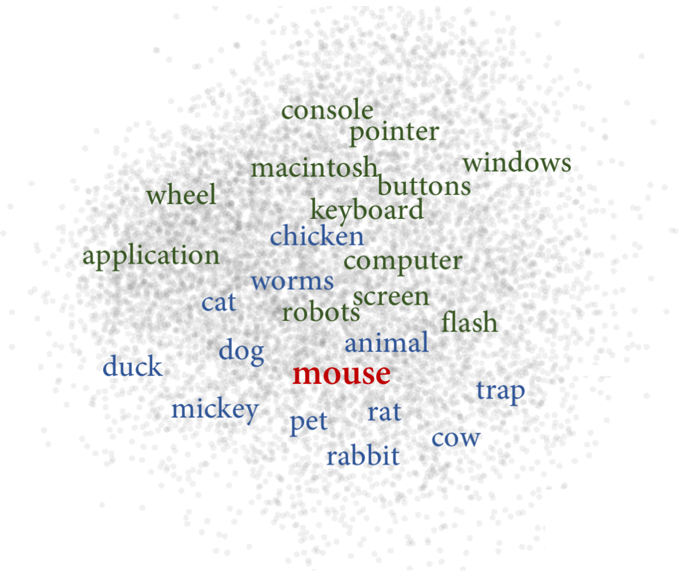

Figure 1: An illustration of the meaning conflation deficiency in a 2D semantic space around the ambiguous word *mouse*. Having the word, with its different meanings, represented as a single point (vector) results in pulling together of semantically unrelated words, such as *computer* and *rabbit*.

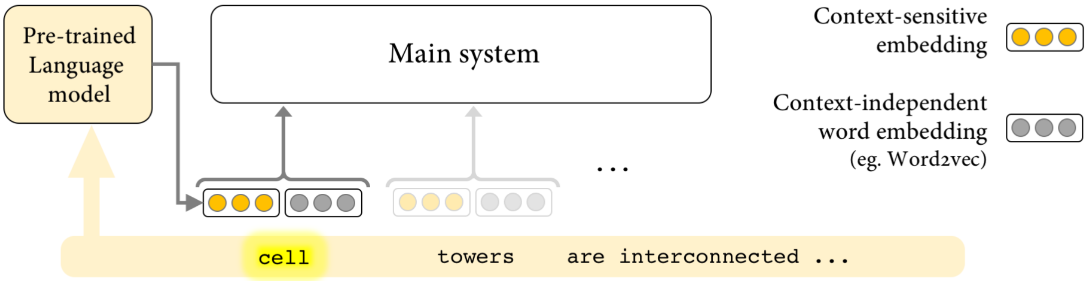
Pre-trained Language model
Main system
...
Context-sensitive embedding
Context-independent word embedding (eg. Word2vec)
cell
towers
are interconnected ...

Figure 2: A general illustration of contextualized word embeddings and how they are integrated in NLP models. A language modelling component is responsible for analyzing the context of the target word (cell in the figure) and generating its dynamic embedding.

i.e., their representations dynamically change depending on the context in which a word appears. Context2vec (Melamud et al., 2016) and ELMo (Peters et al., 2018a) are some of the early examples for this type of representation. These models represent the context of a target word by extracting the embedding of a word in context from a bi-directional LSTM language model. The latter further proposed a seamless integration into neural NLP systems, as depicted in Figure 2. More recently, Transformers (Vaswani et al., 2017) have proven very effective for encoding contextualized knowledge, thanks to their self-attention mechanism (Figure 3). BERT (Devlin et al., 2018), which is based on Transformers, has revolutionized the field of representation learning and has impacted many other fields in NLP. Many derivatives and subsequent models have followed up, rapidly pushing up the state of the art in different benchmarks. In this tutorial we extensively cover this recent type of representation.

We also discuss other types of embeddings, for instance for graph structures which are a popular choice in many scenarios, or for longer units of texts such as sentences and documents. Finally, we conclude this tutorial by discussing some of the ethical issues around the implicit gender and stereotypical biases encoded in word embeddings and proposals for reducing these artifacts.

2 Type of tutorial

Cutting-edge, although the first part of the tutorial could also be considered introductory. The tutorial provides an overview starting from vector space models and word representations and the move to newer

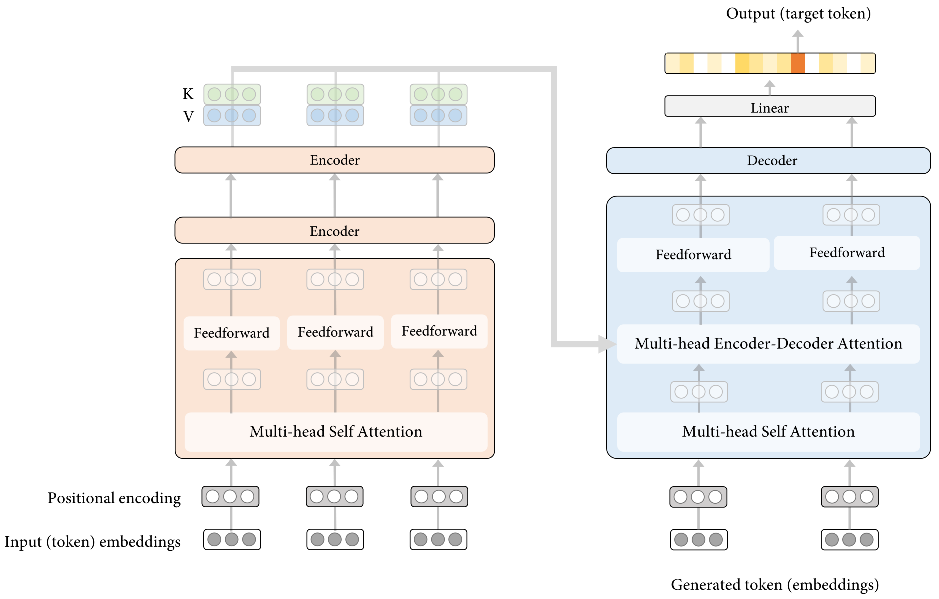

Figure 3: The overall structure of a Transformer-based encoder-decoder model. Many of the recent contextualised models, such as BERT, are based on Transformers.

contextualized embeddings.

3 Outline

The tutorial is split into seven sections, each of which being roughly self-contained.

1. Introduction (20 minutes) A quick warm-up introduction to NLP and why it is important for NLP systems to have a semantic comprehension of texts and how this is usually achieved by representing semantics through mathematical or computer-interpretable notations. This section provides necessary motivation for the tutorial and highlights the role of semantic representation as a core component of many NLP systems. It also briefly describes the Vector Space Model and briefly discusses the evolution path of semantic representations in NLP.

2. Word embeddings (25 minutes) This section explains the main approaches to learn word embeddings from text corpora, what their advantages are and how they have revolutionized the field of lexical semantics. We describe the concepts behind some of the major word embedding techniques, such as Word2vec and GloVe, and their application in NLP. Finally, we briefly cover other types of word embeddings, such as character-based, cross-lingual or knowledge-enhanced.

3. Graph Embeddings (20 minutes) Graphs are ubiquitous data structures. They are often the preferred choice for representing various type of data, including social networks, word co-occurrence and semantic networks, citation networks, telecommunication networks, molecular graph structures and biological networks. In this part of the tutorial, we discuss some of the prominent techniques for transforming graph nodes and edges into vectors. The goal is for the resulting embedding space to preserve the structural properties of the graph, be it the relative positioning of the nodes or relations (edges) among them.

4. Sense Embeddings (20 minutes) In this section we cover those approaches that learn distinct representations for individual meanings of words (i.e., word senses) with the aim of addressing the meaning conflation deficiency. For this part, we will discuss both knowledge-based and unsupervised paradigms.

5. Contextualized Representations (45 minutes) In this part of the tutorial we will introduce the latest type of embeddings that aim at providing dynamic representations of words, capable of adapting the representation to syntactic and semantic characteristics of a given context. We will start by discussing the need for contextualization. We then provide a very brief introduction to the architecture and building blocks of the Transformer model. We then start the overview of contextualized models by some of the earliest proposals in this category, i..e, Context2vec and "Embeddings from Language Models" (ELMo). We then discuss the newer and more prominent models based on Transformers. Specifically, we will describe BERT and some of its derivatives and subsequent works, such as XLNet, DistilBERT, GPT-2 and RoBERTa. We will provide an in-depth analysis of this techniques and point out not only their strengths, but also some of the limitations from which they suffer, which can be taken as possible research directions.

6. Sentence and Document Embeddings (15 minutes) This section goes beyond the level of words, and describes how sentences and documents can be encoded into vectorial representations. We cover some of the widely used supervised and unsupervised techniques and discuss the applications and evaluation methods for these representations. Given the tutorial's main focus on word-level representation, this section provides partial coverage but also pointers for further reading.

7. Ethics and bias (10 minutes) In this section we will talk about the implicit bias in vector representations of meaning, with a focus on gender bias and word representations. We will also overview some of the recent techniques for debiasing word embeddings from gender stereotypes and biases.

4 Breadth

The tutorial is largely based on a recent book written by the instructors published by the Synthesis Lectures on Human Language Technologies of Morgan and Claypool, titled "Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning" (Pilehvar and Camacho-Collados, 2020). This book covers all the various techniques on vectors representations of meaning in detail, while in this tutorial we provide an overview of the main ideas, without getting into details. .

5 Prerequisites for the attendees

No special advanced requirements are required for the attendees, but a certain familiarity with linear algebra, natural language processing and machine learning would be desirable.

6 Small reading list

In addition to the main reference book (Pilehvar and Camacho-Collados, 2020), in the following we present some references that may be helpful for understanding the tutorial. Nonetheless, they are not required to read in advance as their main ideas are also covered as part of the tutorial.

- Schütze (1992): Dimensions of meaning
- Turney and Pantel (2010): From Frequency to Meaning: Vector Space Models of Semantics
- Mikolov et al. (2013): Efficient Estimation of Word Representations in Vector Space
- Pennington et al. (2014): GloVe: Global vectors for word representation
- Melamud et al. (2016): Context2vec: Learning Generic Context Embedding with Bidirectional LSTM
- Peters et al. (2018a): Deep contextualized word representations
- Camacho-Collados and Pilehvar (2018): From word to sense embeddings: A survey on vector representations of meaning
- Peters et al. (2018b): Dissecting Contextual Word Embeddings: Architecture and Representation
- Devlin et al. (2018): BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

7 Instructors

Jose Camacho Collados (camachocolladosj@cardiff.ac.uk; http://www.josecamachocollados.com) is a UKRI Future Leaders Fellow and Lecturer at the School of Computer Science and Informatics at Cardiff University (United Kingdom). Previously he was a Google Doctoral Fellow, completed his PhD at Sapienza University of Rome (Italy) and had pre-doctoral experience as a statistical research engineer in France. His background education includes an Erasmus Mundus Masters in Human Language Technology and a 5-year BSc degree in Mathematics (Spain). Jose's main area of expertise is Natural Language Processing (NLP) and in particular computational semantics or, in other words, how to make computers understand language. His research has pivoted around both scientific contributions through regular publications in top AI and NLP venues such as ACL, EMNLP, AAAI or IJCAI; and applications with direct impact in society, with a special focus on social media and multilinguality. He has also organised several international workshops, tutorials and open challenges with hundreds of participants across the world.

Mohammad Taher Pilehvar (mp792@cam.ac.uk, http://pilehvar.github.io) is an Assistant Professor at Tehran Institute for Advanced Studies (TeIAS) and an Affiliated Lecturer at the University of Cambridge. Taher's research lies in lexical semantics, mainly focusing on semantic representation and similarity. In the past, he has co-instructed three tutorials on these topics (EMNLP 2015, ACL 2016, and EACL 2017) and co-organised three SemEval tasks and an EACL workshop on sense representations. He has also co-authored several conference papers (including two ACL best paper nominations, at 2013 and 2017).

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association of Computational Linguistics*, 5(1):135–146.

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

William A. Gale, Kenneth Church, and David Yarowsky. 1992. A method for disambiguating word senses in a corpus. *Computers and the Humanities*, 26:415–439.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of ACL*, pages 873–882, Jeju Island, Korea.

Jiwei Li and Dan Jurafsky. 2015. Do multi-sense embeddings improve natural language understanding? In *Proceedings of EMNLP*, pages 683–693, Lisbon, Portugal.

Massimiliano Mancini, Jose Camacho-Collados, Ignacio Iacobacci, and Roberto Navigli. 2017. Embedding words and senses together via joint knowledge-enhanced training. In *Proceedings of CoNLL*, pages 100–111, Vancouver, Canada.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of EMNLP*, pages 1059–1069, Doha, Qatar.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Matthew Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018a. Deep contextualized word representations. In *Proceedings of NAACL*, New Orleans, LA, USA.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium, October-November. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2020. Embeddings in natural language processing: Theory and advances in vector representations of meaning. *Synthesis Lectures on Human Language Technologies*, 13(4):1–175.

Mohammad Taher Pilehvar and Nigel Collier. 2016. De-conflated semantic representations. In *Proceedings of EMNLP*, pages 1680–1690, Austin, TX.

Joseph Reisinger and Raymond J. Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Proceedings of ACL*, pages 109–117.

Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of ACL*, pages 1793–1803, Beijing, China.

Hinrich Schütze. 1992. Dimensions of meaning. In *Supercomputing '92: Proceedings of the 1992 ACM/IEEE conference on Supercomputing*, pages 787–796, Los Alamitos, CA, USA.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Amos Tversky and Itamar Gati. 1982. Similarity, separability, and the triangle inequality. *Psychological Review*, 89(2):123.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

George K. Zipf. 1949. *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley, Cambridge, MA.