Beyond Words: Deep Learning for Multiword Expressions and Collocations

Valia Kordoni
Humboldt-Universität zu Berlin
kordonie@anglistik.hu-berlin.de

1 Tutorial Overview

Deep learning has recently shown much promise for NLP applications. Traditionally, in most NLP approaches, documents or sentences are represented by a sparse bag-of-words representation. There is now a lot of work which goes beyond this by adopting a distributed representation of words, by constructing a so-called "neural embedding" or vector space representation of each word or document. The aim of this tutorial is to go beyond the learning of word vectors and present methods for learning vector representations for Multiword Expressions and bilingual phrase pairs, all of which are useful for various NLP applications.

This tutorial aims to provide attendees with a clear notion of the linguistic and distributional characteristics of *multiword expressions* (MWEs), their relevance for the intersection of deep learning and natural language processing, what methods and resources are available to support their use, and what more could be done in the future. Our target audience are researchers and practitioners in machine learning, parsing (syntactic and semantic) and language technology, not necessarily experts in MWEs, who are interested in tasks that involve or could benefit from considering MWEs as a pervasive phenomenon in human language and communication.

This tutorial consists of four parts. Part I starts with a thorough introduction to different types of MWEs and collocations, their linguistic dimensions (idiomaticity, syntactic and semantic fixedness, specificity, etc.), as well as their statistical characteristics (variability, recurrence, association, etc.). This part concludes with an overview of linguistic and psycholinguistic theories of MWEs to date.

For MWEs to be useful for language technology, they must be recognisable automatically. Hence, Part II surveys computational approaches to MWEs recognition, both manually-authored approaches and machine learning ones, as well as computational approaches to MWE elements combination. We will also review type and token evaluation methods for MWE identification.

Part III offers a thorough overview of how and where MWEs can contribute to the intersection of NLP and Deep Learning, particularly focusing on recent advances in the computational treatment of MWEs in the framework of Deep Learning.

Part IV of the tutorial concludes with concrete examples of where MWEs treatment can contribute to language technology applications such as machine translation, information extraction, information retrieval and parsing, as well as MWE-related multi-level annotation platforms (for instance, pipelines) and resources made available for a wide range of languages.

2 Tutorial Outline

1. PART I – General overview:
   (a) Introduction
   (b) Types and examples of MWEs and collocations
   (c) Linguistic dimensions of MWEs: idiomaticity, syntactic and semantic fixedness, specificity, etc.
   (d) Statistical dimensions of MWEs: variability, recurrence, association, etc.
   (e) Linguistic and psycholinguistic theories of MWEs

2. PART II – Computational methods (symbolic and statistical)
   (a) Recognizing the elements of MWEs
   (b) Recognising how elements are combined

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-Tutorial Abstracts, pages 15–16
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-5005

    (c) Type and token evaluation of MWE identification
    (d) Robust automated natural language processing with MWEs

3. PART III – At the intersection of Deep learning and NLP
    (a) Beyond learning word vectors
    (b) Recursive Neural Networks for parsing MWEs
    (c) Learning vector representations for Multiword Expressions, grammatical relations, and bilingual phrase pairs, all of which are useful for various NLP applications

4. PART IV – Resources and applications:
    (a) MWEs in resources: corpora, lexicons and ontologies (e.g., WordNet and Genia), parsers and tools (e.g., NSP, mwetoolkit, UCS, and jMWE), and MWE website (`http://multiword.sf.net`)
    (b) Pipelines for MWE treatment: creation and annotation of resources, identification of MWEs in text, evaluation of results
    (c) MWes in Language Technology applications: Information Retrieval, Information Extraction, Machine Translation

3 Tutorial Instructor

Valia Kordoni is a research professor of computational linguistics at Humboldt University Berlin. She is a leader in EU-funded research in Machine Translation, Computational Semantics, and Machine Learning. She has organized conferences and workshops dedicated to research on MWEs, recently including the EACL 2014 *10th Workshop on Multiword Expressions (MWE 2014)* in Gothenburg, Sweden, the NAACL 2015 *11th Workshop on Multiword Expressions* in Denver, Colorado, and the ACL 2016 *12th Workshop on Multiword Expressions* in Berlin, Germany, among others. She has been the Local Chair of *ACL 2016 - The 54th Annual Meeting of the Association for Computational Linguistics* which took place at the Humboldt University Berlin in August 2016. She has taught a tutorial on *Robust Automated Natural Language Processing with Multiword Expressions and Collocations* in ACL 2013, as well as a tutorial on *Robust Semantic Analysis of Multiword Expressions with FrameNet* in EMNLP 2015, together with Miriam R. L. Petruck. She is also the author of *Multiword Expressions - From Linguistic Analysis to Language Technology Applications* (to appear, Springer).