Machine Learning Applied in Natural Language Processing

Andrei-Mădălin Butnaru
University of Bucharest
butnaruandreimadalin@gmail.com

Abstract

Machine Learning is present in our lives now more than ever. One of the most researched areas in machine learning is focused on creating systems that are able to understand natural language. Natural language processing is a broad domain, having a vast number of applications with a significant impact in society. In our current era, we rely on tools that can ease our lives. We can search through thousands of documents to find something that we need, but this can take a lot of time. Having a system that can understand a simple query and return only relevant documents is more efficient. Although current approaches are well capable of understanding natural language, there is still space for improvement.

This thesis studies multiple natural language processing tasks, presenting approaches on applications such as information retrieval, polarity detection, dialect identification [Butnaru and Ionescu, 2018], automatic essay scoring [Cozma et al., 2018], and methods that can help other systems to understand documents better. Part of the described approaches from this thesis are employing kernel methods, especially string kernels. A method based on string kernels that can determine in what dialect a document is written is presented in this thesis. The approach is treating texts at the character level, extracting features in the form of p-grams of characters, and combining several kernels, including presence bits kernel and intersection kernel. Kernel methods are also presented as a solution for defining the complexity of a specific word. By combining multiple low-level features and high-level semantic features, the approach can find if a non-native speaker of a language can see a word as complicated or not. With one focus on string kernels, this thesis proposes two transductive methods that can improve the results obtained by employing string kernels. One approach suggests using the pairwise string kernel similarities between samples from the training and test sets as features. The other method defines a simple self-training algorithm composed of two iterations. As usual, a classifier is trained over the training data, then is it used to predict the labels of the test samples. In the second iteration, the algorithm adds a predefined number of test samples to the training set for another round of training. These two transductive methods work by adapting the learning method to the test set.

A novel cross-dialectal corpus is shown in this thesis. The Moldavian versus Romanian Corpus (MOROCO) [Butnaru and Ionescu, 2019a] contains over 30.000 samples collected from the news

domain, split across six categories. Several studies can be employed over this corpus such as binary classification between Romanian and Moldavian samples, intra-dialect multi-class categorization by topic, and cross-dialect multi-class classification by topic. Two baseline approaches are presented for this collection of texts. One method is based on a simple string kernel model. The second approach consists of a character-level deep neural network, which includes several Squeeze-and-Excitation Blocks (SE-blocks). As known at this moment, this is the first time when a SE-block is employed in a natural language processing context. This thesis also presents a method for German Dialect Identification composed on a voting scheme that combines a Character-level Convolutional Neural Network, a Long Short-Term Memory Network, and a model based on String Kernels.

Word sense disambiguation is still one of the challenges of the NLP domain. In this context, this thesis tackles this challenge and presents a novel disambiguation algorithm, known as ShowtgunWSD [Butnaru and Ionescu, 2019b]. By treating the global disambiguation problem as multiple local disambiguation problems, ShotgunWSD is capable of determining the sense of the words in an unsupervised and deterministic way, using WordNet as a resource. For this method to work, three functions that can compute the similarity between two words senses are defined. The disambiguation algorithm works as follows. The document is split into multiple windows of words of a specific size for each window. After that, a brute-force algorithm that computes every combination of senses for each word within that window is employed. For every window combination, a score is calculated using one of the three similarity functions. The last step merges the windows using a prefix and suffix matching to form more significant and relevant windows. In the end, the formed windows are ranked by the length and score, and the top ones, based on a voting scheme, will determine the sense for each word.

Documents can contain a variable number of words, therefore employing them in machine learning may be hard at times. This thesis presents two novel approaches [Ionescu and Butnaru, 2019] that can represent documents using a finite number of features. Both methods are inspired by computer vision, and they work by first transforming the words within documents to a word representation, such as word2vec. Having words represented in this way, a k-means clustering algorithm can be applied over the words. The centroids of the formed clusters are gathered into a vocabulary. Each word from a document is then represented by the closest centroid from the previously formed vocabulary. To this point, both methods share the same steps. One approach is designed to compute the final representation of a document by calculating the frequency of each centroid found inside it. This method is named Bag of Super Word Embeddings (BOSWE) because each centroid can be viewed as a super word. The second approach presented in this thesis, known as Vector of Locally-Aggregated Word Embeddings (VLAWE), computes the document representation by accumulating the differences between each centroid and each word vector associated with the respective centroid. This thesis also describes a new way to score essays automatically by combining a low-level string kernel model with a high-level semantic feature representation, namely the BOSWE representation.

The methods described in this thesis exhibit state-of-the-art performance levels over multiple tasks. One fact to support this claim is that the string kernel method employed for Arabic Dialect Identification obtained the first place, two years in a row at the Fourth and Fifth Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial). The same string kernel model obtained the fifth place at the German Dialect Identification Closed Shared Task at VarDial Workshop of EACL 2017. Second of all, the Complex Word Identification model scored a third-place at the

CWI Shared Task of the BEA-13 of NAACL 2018. Third of all, it is worth to mention that the ShotgunWSD algorithm surpassed the MCS baseline on several datasets. Lastly, the model that combines string kernel and bag of super word embeddings obtained state-of-the-art performance over the Automated Student Assessment Prize dataset.

Awarded by: University of Bucharest
Supervised by: Radu Tudor Ionescu
Available at: bit.ly/Butnaru-PhD-Thesis

Selected Publications

Andrei M. Butnaru and Radu Tudor Ionescu. UnibucKernel Reloaded: First Place in Arabic Dialect Identification for the Second Year in a Row. In *Proceedings of VarDial Workshop of COLING*, pages 77–87, 2018.

Mădălina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. Automated essay scoring with string kernels and word embeddings. In *Proceedings of ACL*, pages 503–509, 2018.

Andrei M Butnaru and Radu Tudor Ionescu. MOROCO: The Moldavian and Romanian Dialectal Corpus. *In Proceedings of Association of Computational Linguistics*, 2019a.

Andrei M Butnaru and Radu Tudor Ionescu. Shotgunwsd 2.0: An improved algorithm for global word sense disambiguation. *IEEE Access*, 7:120961–120975, 2019b.

Radu Tudor Ionescu and Andrei Butnaru. Vector of locally-aggregated word embeddings (vlawe): A novel document-level representation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 363–369, 2019.