Abstract

Keywords:

1. Outline

- Intro
  - word embedding ...
  - history, general approaches
  - weakness of previous approach
    - unsupervised approaches
    - lexical resource based approaches (Bordes)
  - explain why we want to use lexical resource (advantages of using lexical resources)
  - explain availability of datasets
  - explain possible applications (combining word embedding)
- Related work
  - word embedding
  - Lexical resources
- word embedding approach
  - Bordes word embedding (AAAI 2011)
  - Uby
    - combining different resources
    - statistics
    - easy access (API)
  -

- application in NLP tasks (SENNA extension)
- experiment discussion
  - experimental setting
  - Evaluation measure
  - evaluation result
  - interpretation (improvement, limitation, ...)
- conclusion

2. Related Work

In this chapter, we will define and justify the task of *Representation Learning* and we will see different families of methods for inducing word representation and its application in NLP. In machine learning specially in industry, most of the labor is dedicated to *Feature Engineering*. Extracting informative features is the crucial part of most supervised methods and it is done mostly manually. While many different applications share common learning models and classifiers, the difference in performance of competing methods mostly goes to the data representation and hand-crafted features that they use. This observation reveals an important weakness in current models, namely their inability to extract and organize discriminative features from data. Representation learning is an umberella term for a family of unsupervised methods to learn features from data. Most of recent works on the application of this idea in NLP focus on inducing word representations. *Word representation* is a mathematical object, usually a vector, which each dimension in this vector represents a grammatical or semantical feature to identify this word and is induced automatically from data [1]. Recently, it has been shown in [1] and [2] that using induced feautres can be helpful to improve state-of-the-art methods in different NLP tasks. In section **??** we will describe one possible way of incorporating this idea in to our task. In the next two sections, two major families of representations will be shortly reviewed.

2.1. Distributional Representation

In distributional semantics, the meaning of a word is expressed by the context that it appears in it [3]. Features that are used to represent the

meaning of a word are other words in its neighborhood as it is so called the context. In some approaches like LDA and latent semantic analysis (LSA), the context is defined in the scope of a document rather than a window around a word. To represent word meanings in via distributional approach, one should start from count matrix (or zero-one co-occurence matrix) which each row represents a word and each column is a context. The representation can be limited to raw usage of the very same matrix or some transforms like *tf-idf* will be applied first. A further analysis over this matrix to extract more meaningful features is applying dimensionality reduction methods or clustering models to induce latent distributional representations. A similar clustering method to k-means is used in [4] to represent phrase and word meanings and brown clustering algorithm [5] has been shown to have impact on near to state-of-the-art NLP tasks [1].

2.2. Distributed Representation

Distributed representation has been introduced in the literature for the first time in [6] where Bengio et al. introduced a first language model based on deep learning methods[7]. Deep learning is learning through several layers of neural networks which each layer is responsible to learn a different concept and each concecpt is built over other more abstract concepts. In the deep learning society, any word representation that is induced with a neural network is called *Word Embedding*. In contrast to raw count matrix in distributional representations, word embeddings are low-dimensional, dense and real-valued vectors. The term, **'Distributed'**, in this context refers to the fact that exponential number of objects (clusters) can be modeled by word embeddings. Here we will see two famous models to induce for such representations. One family will use n-grams to learn word representation jointly with a language model and the other family learns the embedding from structured resources.

2.3. Neural Language Models

In [8], Weston and Collobert use a non-probabilistic and discriminative model to jointly learn word embeddings and a language model that can separate plausible n-grams from noisy ones. For each word in a n-gram, they combine the word embeddings and use it as positive example. They put noise in the n-gram to make negative examples and then train a neural network to learn to classify postive labels from negative ones. The parameters of neural

network (neural language model) and word embedding values will be learned jointly by an optimization method called *Stochastic Gradient Descent* [9].

A hierarchical dsitributed language model (HLBL) proposed by Mnih and Hinton in [10] is another influential work on word embeddings. In this model a probabilistic linear neural network(LBL) will be trained to combine word embeddings in first $n-1$ words of a n-gram to predict the n_th word.

Weston-Collobert model and HLBL by Mnih and Hinton are evaluated in [1] in two NLP tasks: chunking and named entity recognition. With using word embeddings from these models combined with hand-crafted features, the performance of both tasks are shown to be improved.

2.4. Representation Learning from Knowledge Bases

Bordes et al. in [11] and [12] have attempted to use deep learning to induce word representations from lexical resources such as WordNet and knowledge bases (KB) like Freebase. In Freebase for example, each named entity is related to another entity by an instance of a specific type of relation. In [11], each entity is represented as a vector and each relation is decomposed to two matrices. Each of these matrices transform left and right-hand-side entities to a semantic space. Similarity of transformed entities indicates that the relation holds between the entities. A prediction task is defined to evaluate the embeddings. Given a relation and one of the entities, the task is to predict the missing entity. The high accuracy (99.2%) of the model on prediciton of training data shows that learnt representation highly captures attributes of the entities and relations in Freebase.

References

[1] J. Turian, L. Ratinov, Y. Bengio, Word representations: a simple and general method for semi-supervised learning, ACL (2010) 384–394.

[2] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural Language Processing (almost) from Scratch, Machine Learning Research 12 (2011) 2493–2537.

[3] Z. Harris, Distributional structure, Springer Netherlands, 1981. URL: http://link.springer.com/chapter/10.1007/978-94-009-8467-7_1.

[4] D. Lin, X. Wu, Phrase clustering for discriminative learning, in: ACL-AFNLP, 2009, pp. 1030–1038. URL: http://dl.acm.org/citation.cfm?id=1690290.

[5] P. Brown, P. Desouza, R. Mercer, Class-based n-gram models of natural language, Computational Linguistics 4 (1992) 467–479.

[6] Y. Bengio, R. Ducharme, A neural probabilistic language model, The Journal of Machine ... 3 (2003) 1137–1155.

[7] Y. Bengio, Learning deep architectures for AI, Foundations and Trends in Machine Learning (2009).

[8] R. Collobert, J. Weston, A unified architecture for natural language processing: deep neural networks with multitask learning, ... international conference on Machine learning (2008).

[9] L. Bottou, Large-scale machine learning with stochastic gradient descent, Proceedings of COMPSTAT'2010 (2010).

[10] A. Mnih, G. Hinton, A scalable hierarchical distributed language model, Advances in neural information processing systems (2009) 1–8.

[11] A. Bordes, J. Weston, R. Collobert, Y. Bengio, Learning structured embeddings of knowledge bases, AAAI (2011) 301–306.

[12] A. Bordes, X. Glorot, J. Weston, Y. Bengio, Joint learning of words and meaning representations for open-text semantic parsing, in: Artificial Intelligence and Statistics (AISTATS), volume 22, 2012. URL: http://redirect.subscribe.ru/_/-/jmlr.csail.mit.edu/proceedings/papers/v22/