Learning Word Representations from a Large-Scale Unified Lexical Semantic Resource

Abstract

Learning word representations and iducing word feautres are shown to be able to improve the performance in various NLP tasks such as Word Sense Disambiguation, Named Entity Recognition, Parsing,...

Here in this paper, we investigate the effectivness of learned features for words from structured knowledge bases with focus on a method proposed by Bordes et al. We extend their idea with incorporating multiple resources from different languages (English and German) and also different type of resources (WordNet, FrameNet). We have evaluated both monolingual (Bordes embeddings) and bilingual embeddings (our embeddings) on four different gold dataset for word-pair similarity task and shown that bilingual embeddings perform similarly or better than monolingual embeddings.

Keywords: Representation Learning, Word Embeddings, Machine Learning, Semantics

1. Introdcution

In a large number of machine learning methods and its application to natural language processing, 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* or *Word embedding* "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 word representations can be helpful to improve state-of-the-art methods in variouse NLP tasks. In Section 2, some of these methods are discussed in more details.

From recent works, we observe that most of the current methods for inducing word representations can only exploit surface relation among words. Indeed, the only resource for them to capture semanitcal and grammatical aspects of words is co-occuring of them in a text. The word embeddings learned in neural language models ([3] and [4]) and brown clustering are examples of such approach. In the contrast to these methods, Bordes et al. [5] proposed a method to learn distributed representations from relational datasets with richer information. In their work, they are attempting to induce word embeddings from knowledge bases such as WordNet and Freebase. Their datasets include binary relations between left entity and right entity and each relation is instantiated from a different relation type. Since we are following their methodology, a detailed description of their work is presented in 2.4.

After reviewing previous related works, we will demonstrate our contribution for inducing word embeddings from multiple lexical resources and show its effectiveness for inducing bilingual word embeddings and transfering information from one language to another one. A pipleline of our system for combining different lexical resources to capture broader grammatical and semantical features than previous works into our word embeddings will be described in detail. ??? Uby as a unified lexical resource which plays a central role in our system will be reviewd shortly ???

Finally, we will evaluate our word embeddings empirically in different settings as a proof-of-concept to show the role of representation learning jointly from multiple lexical resources. We will also zoom in to our learned embeddings for special case of English-German to inspect the strength of bilingual word embeddings.

(??? Parsing with Compositional Vector Grammars Socher et al. ACL 2013, . Improving Word Representations via Global Context and Multiple Word Prototypes Huang 2012 ???)

2. Related Work

2.1. Distributional Representation

In distributional semantics, the meaning of a word is expressed by the context that it appears in it [6]. 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 [7] to represent phrase and word meanings and brown clustering algorithm [8] 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 [4] where Bengio et al. introduced a first language model based on deep learning methods[9]. 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. (Cross-lingual Word Clusters for Direct Transfer of Linguistic Structure should be mentioned ???)

2.3. Neural Language Models

In [3], 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* [10].

A hierarchical dsitributed language model (HLBL) proposed by Mnih and Hinton in [11] 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

???(should be expanded with mathematical notation and better description of their models and experiments)??? Bordes et al. in [5] 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 [5], 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.

3. Our contribution

3.1. Uby

3.2. Bilingual word embeddings

???(transfer learning and multi task learning should be mentioned from Caruana, R. (1997). Multitask Learning. Machine Learn- ing, 28, 4175. Chapelle,)???

As it is described in the previous section we can relate two senses from two different resources using Uby SenseAxis Alignments. This is an additional information which can play a role of bridge between two different datasets to transfer knowledge from one to the another. Using this new feature we make our WordNet-GermaNet dataset which contains three type of relations (1) WordNet relations (2) GermaNet relations (3) Cross-lingual sense relations between WordNet and GermaNet
Example of relations:

WN-1	rel1	WN-2
GN-1	rel3	GN-2
WN-1	c-rel	GN-2

We have also created another version of this dataset but with different granularities, we mapped similar inter-lingual relations to same relations. This will help to have faster learning phase with roughly similar performance.

Since cross-lingual sense alignments are expressing nearly-synonym relation among two senses and the Bordes model is sensitive to the direction of relations we have added the reverse sense alignments too to encode bidirectional nature of this type of relations.

infering wordnet-framenet data. WordNet and GermaNet are expressing similar knowledge but in different languages, so it is worthwile to examine learning word embeddings from two different knowledge base which contains different semantical aspects of words and their senses. Therefor by using Uby and the method described in [CM FN-Wkt] we infered relations among WordNet and FrameNet. FrameNet is blah blah..

In the next section, we will describe the different settings to analyze performance of learned embeddings from our new datasets.

4. Empirical Evaluation

To show the effectiveness of joint learning of features from multiple knowledge bases we suggest two experiment setups. In the first schema we follow Bordes et al. ranking task. The goal of this task is to show how good the structure of knowledge bases are represented through the learned features. After we learned the word embeddings from subset(??) of Uby(??), their ability to reproduce the structure of it will be assessed. On the other hand, the second setup is investigating on this question that if the learned word embeddings from multiple resources are able to improve the performance of original Bordes model in a standard NLP task, here word-pair similarity or not. In this setup we will look to contribution of the learned features in predicting similarity of words.

4.1. Intrinstic Evaluation

Bordes et al. define a ranking task where for each triplet (e_l, r, e_r) in trianing and test set, e_l will be removed and all the entities will be ranked by using 1-norm rank function (equation ??? decomposing equation). A higher rank of e_l (lower number) reflects the better quality of learned representations. Additionally they have compared this result to another ranking schema using density estimation . In this schema, for each word embedding e the density of (e, r, e_r) will be computed (as it is described in our section???) and triplets will be sorted by their estimated probability (probability terms ??). Since we are using larger sets of triplets, instead of ranking all the training instances we sample randomly from each training dataset with size of 20% of the original dataset(??) then we test our models on these sampled training instances and all the instances from test set. Bordes et al. have followed a similar approach for ranking their embeddings on their biggest dataset. We re-run their related experiments to make the comparison to our embeddings meaningful. Table (??) shows the results.

We repeat the ranking evaluation with two different embeddings: (1) learned from GermaNet (2) jointly learned from GermaNet-WordNet. The intrinsic evaluation we use here can't be used to compare the effectivness of these two different embeddings since the evaluation only reflects the difficulty level of a structure and since these

Table (??) presents the comparison of ranking tasks for mono-lingual and bilingual word embeddings.

Table 1: Ranking Performance for Non-mapped Relations

Dataset	#dimension	#relations	#entities		Micro	Macro
GermaNet	25	16	64025	lhs	82.08	73.11
				rhs	81.22	72.36
				global	81.65	72.74
WordNet	25	23	148976	lhs	81.76	85.79
				rhs	81.96	85.49
				global	81.86	85.63
WordNet-GermaNet (WN)	25	32	213002	lhs	82.50	85.09
				rhs	83.16	84.46
				global	82.83	84.78
WordNet-GermaNet (GN)	25	32	213002	lhs	72.12	63.63
				rhs	67.78	65.77
				global	69.95	64.70
WordNet-FrameNet	25	25	25	lhs	1	1
				rhs	1	1
				global	1	1

Table 2: Ranking Performance for Mapped Relations

Dataset	#dimension	#relations	#entities		Micro(%)	Macro
GermaNet	25	10	64025	lhs	82.60	68.1
				rhs	81.90	68.8
				global	82.25	68.5
WordNet	25	19	148976	lhs	83.50	83.1
				rhs	84.22	83.6
				global	83.86	83.4
WordNet-GermaNet (WN)	25	24	213002	lhs	78.70	82.6
				rhs	79.56	83.0
				global	79.13	82.8
WordNet-GermaNet (GN)	25	24	213002	lhs	69.66	59.5
				rhs	66.60	58.9
				global	68.13	59.2
WordNet-FrameNet	25	25	25	lhs	1	1
				rhs	1	1
				global	1	1

4.2. Extrinsic Evaluation

We are interested to further analyze the role of multi-task learning of embeddings for transforming knowledge from one resource to the another. In order to examine if semantic information from English (WordNet) can be transfered to German (GermaNet) or the other way, we compare the embeddings learnt from multiple resources to the embeddings learnt from single resource in word-pair similarity experiments. Four datasets of word-pair similarity are used to compare the correlation of predicted similairty of pair of words against human judgments. [rubensteinGoodenough], [yang-Powers], [millerCharles] and [finkelstein] are datasets that we used to meaure the correlation of similarities predicted by the original bordes model (single resource) and our proposed model (multiple resource) to human judgments. To measure the similarity between any given wordpair (w_1, w_2) we find all vectors associated to different senses of the given words in our embedding dictionary and compute and find the maximum cosine similarity between two vectors. Then for each dataset, both Pearson and Spearman correlation among predicted and gold similarities were calculated which is reported in table 3.

Table 3: Word-pair Similarity Performance for English

Dataset		WN-SE50	WN-GN-SE50	WN-SME-BIL50	WN-
RubensteinGoodenough65	Pearson	0.488	0.571	00	
	Spearman	0.426	0.528	00	
MillerCharles30	Pearson	0.454	0.438	00	
	Spearman	0.40	0.34	00	
Finkelstein353	Pearson	0.194	0.177	00	
	Spearman	0.137	0.128	00	
YangPowers130	Pearson	0.634	0.771	00	
	Spearman	0.598	0.770	00	

As we see in the table 3 in two datasets the performance of learned embeddings from bi-lingual resources are slightly worse but comparable to the mono-lingual embeddings and in the other two datasets one can observe a significant increase of performance of bi-lingual resources over monolingual resources.

5. Conclusion and Future Work

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

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

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

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

[7] 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`.

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

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

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

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

[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/