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

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

Word representations and specially word feautres induced by distributed models are shown to be able to boost the performance of various NLP tasks such as Word Sense Disambiguation, Named Entity Recognition, Parsing,... Bordes et al. have proposed a model in [??2011] to learn representation for entities of a structured knowldege base such as WordNet.

Here in this paper, we follow [??2011] and extend their idea by incorporating multiple resources unified and linked through UBY in order to induce richer representations jointly for two different languages. 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.

1 Introduction

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.*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" (**?**). Recently, it has been shown in (**?**) and (**?**) that using induced word representations can be helpful to improve state-of-the-art methods in variouse NLP tasks. While their word embeddings are induced for a single language, Klementiev et al. [Inducing Crosslingual Distributed Representations of Words] have a model which learns cross-lingual representations and is shown to have superior performance for text classification task. In contrast to previous similar works which word embeddings learnt from a corpus, Bordes et al. proposed a method (**?**) to learn distributed representations from multi-relational knowledge bases such as WordNet and Freebase. Their datasets include binary relations between two entities and each relation is instantiated from a different relation type. Since we are following their methodology, a description of their work is presented in **??**.

In the rest of paper we introduce our method to extend their idea to learn bi-lingual word embeddings from multiple resources. Our contribution is to 1)infere cross-resource and cross-lingual relations has been enabled us to share the task of learning embeddings between different resources and languages and 2) encoding this information in a way that it can be fed to Bordes model. ... Uby is something sime hting is providing a necessary described in **??**.

Finally we will compare performance of monolingual and bi-lingual embeddings in word similarity task to investigate the effectiveness of them to capture different aspects and features of words meanings(??).

2 Representation Learning from Knowledge Bases

Manuscripts must be in two-column format. Exceptions to the two-column format include the title, authors' names and complete addresses, which must be centered at the top of the first page, and any full-width figures or tables (see the guidelines in Subsection **??**). **Type single-spaced.** Start all pages directly under the top margin. See the guidelines later regarding formatting the first page (Subsection **??**). The manuscript should be printed single-sided and its length should not exceed the maximum page limit described in Section **??**. **Do not number the pages.**

2.1 Bordes model for word embedding

Bordes et al. in (**?**) and (**?**) have attempted to use distributed models to induce word representations from lexical resources such as WordNet and knowledge bases (KB) like Freebase. Their model in [2011] is composed of a two major elements.

In WordNet for example, each synset is related to another synset by an instance of a specific type of relation. In (**?**), 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.

2.2 Uby

this part is dedicated to introudce Uby and Uby API.

2.3 Creating of Dataset

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 tasks 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. For example , in this encoding, [list of relatons] are mapped to [rel1].

We will compare them later together to examine the sensitivity of model to different granularities of relations.

Some statistics of data should be shown here.

3 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 monolingual embeddings 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.

3.1 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 dif-

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

Datasets	#dimension	#relations	#entities
GermaNet	25	10	64025
WordNet	25	19	148976
WordNet-GermaNet (WN)	25	24	213002
WordNet-GermaNet (GN)	25	24	213002
WordNet-FrameNet	25	25	25

ferent 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.

3.2 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], [yangPowers], [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- [cut off]
RubensteinGoodenough65	Pearson	0.488	0.57 [cut off]
	Spearman	0.426	0.52 [cut off]
MillerCharles30	Pearson	0.454	0.43 [cut off]
	Spearman	0.40	0.34 [cut off]
Finkelstein353	Pearson	0.194	0.17 [cut off]
	Spearman	0.137	0.12 [cut off]
YangPowers130	Pearson	0.634	0.77 [cut off]
	Spearman	0.598	0.77 [cut off]

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.

4 Conclusion and Future Work

Papers that had software and/or dataset submitted for the review process should also submit it with the camera-ready paper. Besides, the software and/or dataset should not be anonymous.

Please note that the publications of EACL-2014 will be publicly available at ACL Anthology (http://aclweb.org/anthology-new/) on April 19th, 2014, one week before the start of the conference. Since some of the authors may have plans to file patents related to their papers in the conference, we are reminding authors that April 19th, 2014 may be considered to be the official publication date, instead of the opening day of the conference.

Acknowledgments

Do not number the acknowledgment section. Do not include this section when submitting your paper for review.

References