Learning Bi-lingual Word Representations using Uby 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,... Previously a model have been proposed to learn representation for entities of a structured knowldege base such as WordNet.

Here in this paper, we follow the previous work and extend their idea by incorporating multiple resources in order to induce richer representations jointly for two different languages. We have evaluated both monolingual and bilingual embeddings on four different gold dataset for word-pair similarity task and shown that bilingual embeddings perform similarly or better than monolingual embeddings. For example on one of datasets, bilingual embeddings is 10 percent more correlated (Pearson correlation) to human judgements than monolingual embeddings of the same model and up to 40 percent more than the other models.

1 Introduction

In a large number of machine learning methods and their application to computational linguistics *feature engineering* or extracting informative features is a crucial part and it is done mostly manually. *Representation Learning* is an umberella term for a family of unsupervised methods to learn features from data to decrease the manual labour. Most of recent works related to this idea 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 standard 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 over strong baselines. 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(KB) such as WordNet and Freebase. They encode information in KBs as binary relations between entities which each relation is instantiated from a set of relation types. Since we are following their methodology, a description of their work is presented in **??**.

This paper is motivated by two questions. The first question is that how can we extend the framework in (Bordes2011) to induce crosslingual word representations to benefit from aggregation of information in two different resource in two (or more) different languages. The second question is how much informative are these learned features? More specifically, we inspect the degree of ability of these embeddings and also embeddings from other methods to capture semantic similarity between words.

Regarding the first question, In section ?? we introduce our approach to use Bordes framework to learn bi-lingual word embeddings from multiple resources. Our contribution is to 1)infere cross-resource and cross-lingual relations which will enable 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.

Finally, section ?? tries to answer the second

question. Therefor we will compare performance of mono-lingual and bi-lingual embeddings induced by different methods as well as ours 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

In this section, we will first review a framework proposed in (Bordes2011 and Bordes2012) and then will give a short introduction on Uby. In the last part of this section, we will show how to use information encoded in or inferred from Uby to induce cross-lingual word representation with Bordes et al models.

2.1 Bordes model for word embedding

Two major models are proposed in [Bordes2011] and [Bordes2012] to transfer information in Knowledge Bases (KBs) , encoded as graphs, to continues vector space. The knowledge representation in most of KBs can be expressed in form of triples of (e_i, r_k, e_j) where e_i and e_j are i_{th} and j_{th} entities related by a binary relation of type r_k. The purpose of the models is to induce a vector space and associate each entity relation to an embedding vector (or a matrix in the case of first model for relations) which its dimensions are supposed to reflect a set of informative features of entities and relations.

In the first model, **structured embeddings**, entities are modeled as *d*-dimensional vectors. An associated vector to the i_{th} entity, e_i, is $E_i \in \mathbb{R}^d$. Each relation , r, is decomposed to two operators each represented as $d \times d$ matrix, R_{left} and R_{right}.

2.2 Creating of Dataset

Uby is a unified lexical resource which to our approach for providing cross-resource information. For more information on Uby we will point the readers to this (?) and that(?). 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 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.

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 4.

As we see in the table 4 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.

More analysis on why some dataset is good and some is not good.

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.

References

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

Table 2: Ranking Performance for Mapped Relations

Dataset	#dimension	#relations	#entities		Micro(%)	Macro(%)
GermaNet	25	10	64025	lhs	82.60	68.18
				rhs	81.90	68.84
				global	82.25	68.51
WordNet	25	19	148976	lhs	83.50	83.17
				rhs	84.22	83.64
				global	83.86	83.40
WordNet-GermaNet (WN)	25	24	213002	lhs	78.70	82.60
				rhs	79.56	83.06
				global	79.13	82.83
WordNet-GermaNet (GN)	25	24	213002	lhs	69.66	59.54
				rhs	66.60	58.95
				global	68.13	59.25

Table 3: Word-pair Similarity Performance for English

Dataset		WN-SE	WN-GN-SE	WN-SME-BIL	WN-GN-SME-BIL	HLBL	Turian	KlementievTit
RubensteinGoodenough65	P	0.488	0.571	00	00			
	S	0.426	0.528	00	00			
MillerCharles30	P	0.454	0.438	00	00			
	S	0.40	0.34	00	00			
Finkelstein353	P	0.194	0.177	00	00			
	S	0.137	0.128	00	00			
YangPowers130	P	0.634	0.771	00	00			
	S	0.598	0.770	00	00			

Table 4: Word-pair Similarity Performance for German

Dataset		GN-SE	WN-GN-SE	GN-SME-BIL	WN-GN-SME-BIL	KlementievTitov
wortpaare222	P	-0.022	0.112	0.058	00	0.118
	S	-0.100	0.225	0.230	00	0.153
wortpaare30	P	0.865	0.984	-0.443	00	-0.887
	S	1.0	1.0	-0.500	00	-1.0
wortpaare350	P	-0.089	0.045	0.163	00	0.168
	S	-0.158	-0.017	0.135	00	0.117
wortpaare65	P	0.800	0.558	-0.572	00	0.233
	S	0.800	0.800	-0.8	00	0.200