Towards cross-lingual utilization of sparse word representations

Gábor Berend

University of Szeged,
Department of Informatics
2. Árpád tér, 6720, Szeged, Hungary
berendg@inf.u-szeged.hu

Abstract. In this paper, we introduce our approach for defining sparse word representations that are comparable across languages. The proposed approach is designed to require as little linguistic resources as possible. Our experimental results suggests that it is sufficient to rely on an artificially created highly noisy dictionary to map sparse word representations across languages.

1 Introduction

Distributed word representations [1,2] aim at representing symbolic word forms as some relatively low dimensional continuous vectors. Such word representations have been found to be extremely useful in a variety of natural language processing tasks. In fact most state-of-the-art approaches in NLP rely on their utilization [3].

Over the past few years, there has also been an emerging research interest in constructing word representations which are applicable over a variety of languages, see e.g. [4] for an empirical comparison. Providing word representations which are comparable across languages have huge potential because they could enable the effortless creation of machine learning models which can operate across languages. As an illustrative example, if one has access to comparable word representations for both English and French, it becomes possible to train a document classifier purely on English-written documents and then reliably apply it on French documents as well.

A further line of research which has gained attention is related to the application of sparse word representations [5,6]. Such sparse word representations have been shown to often outperform the application of dense word representations in a monolingual setting. However, to the best of our knowledge, there has not been any work conducted towards the creation of sparse word representations which are comparable across languages. In this paper, we present our proposed solution for solving that task by creating and evaluating sparse word representations for English and Hungarian which are intended to be used interchangeably.

2 The proposed solution

In this section we introduce our proposed solution for determining sparse word representations which are comparable across languages. Throughout the rest of the paper we shall denote the dense word embeddings of the source (English) and target (Hungarian) languages by $S \in \mathbb{R}^{k \times |V_s|}$ and $T \in \mathbb{R}^{k \times |V_t|}$, respectively, with V_s and V_t indicating the vocabulary of the two languages. Recall that in the followings we shall denote some symbolic word form as x and the vectorial representation that gets assigned to it in boldface, i.e. $\mathbf{x}$.

2.1 Preprocessing of continuous word embeddings

Our algorithm takes as input a pair of "traditional" dense word embeddings for both the source and target languages trained independently. Pre-trained word embeddings for a variety of languages are already accessible and we relied on the 64-dimensional pre-trained polyglot word embeddings [7]. Utilizing fasttext embeddings [8] would have been a viable alternative. In fact, the proposed approach naturally generalizes to the application of word embeddings trained in any other kinds of ways. We also conducted experiments with dense word embeddings trained according to the popular continuous bag-of-words and skip-gram models but experienced no significant difference in the results.

We performed experiments with the embeddings in their unmodified form as well as experiments that modified the pre-trained word embeddings in ways commonly met in the literature [9,10]. The two kinds of preprocessing steps we performed are making the word embeddings unit length and centralizing the word representations across all the dimensions.

2.2 Mapping of word embeddings

In order to map word representations between English and Hungarian that are trained independently, we perform a linear mapping which brings Hungarian word representations as close to semantically similar English word representations as possible. As firstly proposed in [11] such a linear mapping M can be defined by minimizing the objective function $\sum_{i=1}^{n} \|M\mathbf{s_i} - \mathbf{t_i}\|$, with $\{(s_i, t_i)\}_{i=1}^{n}$ being the seed set of word pairs which are cross-lingual equivalents of each other.

Multiple studies have showed recently that ensuring the additional constraint for M to be orthonormal can significantly improve the quality of the mapping of word embeddings from one language to the embedding space of another one [12,9,10]. Requiring M to be orthonormal can be especially useful if the set of seed word pairs used for determining the linear mapping contains a high fraction of noise, i.e. erroneously aligned word pairs.

Since our goal is to design an approach for mapping sparse word representations which relies on as little human labor and external resources (such as parallel text) as possible, we created a pseudo-dictionary similar to [12]. We aligned word forms that are present in their exact same surface forms in the

vocabulary of both languages. Recall that the pseudo-dictionary created in such a manner undoubtedly contains substantial noise, e.g. the English noun "*hat*" (referring to the clothing accessory) gets aligned to the Hungarian word with the same surface form which – on the other hand – can either refer to a numeral and a verb as well.

The pseudo-dictionary constructed that way consists of 20,292 entries. The full size of the English and Hungarian vocabularies of the polyglot vectors include 100K and 150K word forms. The large number of aligned word pairs relative to the vocabulary sizes also implies that this automatically generated dictionary is not really reliable for which reason the orthonormal constraint for the linear mapping matrix M is expected to provide important gains over its unconstrained counterpart.

2.3 Determining the cross lingual representations

Previous studies have shown that trying to reconstruct dense word embeddings as a sparse linear combination of an overcomplete set of basis vectors can provide useful word representations [5,6]. More formally, for some word embedding matrix $X \in \mathbb{R}^{k \times |V|}$, such approaches seek to find a decomposition for X such that $\|X - D\alpha\|_F + \lambda\|\alpha\|_1$ gets minimized, with $D \in \mathbb{R}^{k \times l}$ containing the set of overcomplete basis vectors and $\alpha \in \mathbb{R}^{l \times |V|}$ containing the sparse linear coefficients for the individual word forms, respectively. In the previous expression λ is the regularization coefficient which controls for the amount of sparsity emerging in the sparse coefficient matrix α.

The kind of decomposition we employ in this work differs both from [5] and [6] as here we not only require D to be a member of the convex set of matrices comprising of unit norm column vectors, but also enforce the coefficients in α to be non-negative. We used the SPAMS package accompanying [13] for performing the matrix decompositions for our experiments.

Putting the various steps together, our approach can be summarized in the following:

1. (Optionally) pre-process the embedding matrices S and T by making the embeddings unit long and centered at the origin
2. Create a (pseudo-)dictionary $\{(s_i, t_i)\}_{i=1}^{n}$
3. Find M for which $\sum_{i=1}^{n} \|M\mathbf{s_i} - \mathbf{t_i}\|$ is minimized (with the optional constraint for M to be orthonormal).
4. Find D_s and α_s such that $\|S - D_s\alpha_s\|_F + \lambda\|\alpha_s\|_1$ gets minimized
5. Find α_t by relying on D_s such that $\|MT - D_s\alpha_t\|_F + \lambda\|\alpha_t\|_1$ is minimized.

Since target word vectors are mapped to the embedding space of source word vectors and the same dictionary matrix D_s is used for decomposing the embedding matrices S and T, the nonzero coefficients in α_s and α_t provide a sparse representation being comparable across languages. Note that if we choose M to be the identity matrix, we can keep the target language embeddings intact.

3 Experiments

During our experiments, we treated English as the source language and Hungarian as the target language. Throughout our experiments, we relied on pre-trained polyglot embeddings [7] which are publicly available for a variety of languages.[1] In order to conduct comparable experiments with previous studies, we applied the same number of basis vectors (i.e. 1024) which was previously utilized in [5]. When setting the value for λ, we chose the regularization coefficient from $\{0.1, 0.3, 0.5\}$ in order to see how different sparsity levels influence results. The sparse word representation that we define for some word form w_i is simply taken by the indices of the positive coefficients in α for the particular word, that is $\phi(\mathbf{w}_i) = \{j \mid \boldsymbol{\alpha}_i[j] > 0\}$, where $\boldsymbol{\alpha}_i$ refers to the vector of sparse coefficients determined for word w_i during the sparse decomposition procedure.

3.1 Evaluation on the Swadesh word list

Swadesh lists are collections of English words of varying cardinalities that are collected as part of a universal basic vocabulary[14]. The elements of the lists are expected to be found in the majority of languages. Consequently, the lists have been translated into various languages, including Hungarian as well.

There exist multiple Swadesh lists with different amount of words included in them. During our experiments we utilized the Swadesh list translated for Hungarian[2] which consists of 207 word forms.

Unlike the pseudo-dictionary that we created for learning the linear mapping between the embedding space of the two languages, the quality of the translation pairs is much higher in this case. We should also note that none of the translation pairs originating from the English and Hungarian Swadesh lists are included in the automatically created pseudo-dictionary.

For the 207-element Swadesh list there are 161 words for which Hungarian translations are given in a one-to-one manner, thus we perform our evaluation over these word pairs alone. During this evaluation phase, we calculate for every word pair (s_i, t_i) – located in the filtered Swadesh list – the extent of overlap in the sparse representations of words s_i and t_i. We define precision and recall as $P = \frac{|\phi(s_i) \cap \phi(t_i)|}{|\phi(t_i)|}$ and $R = \frac{|\phi(s_i) \cap \phi(t_i)|}{|\phi(s_i)|}$, respectively. In order to get an aggregated score for quantifying the overlap between the sparse representation of semantically equivalent words in different languages, we take the harmonic mean (F-score) of the precision and recall scores.

Table 1 includes the results quantifying the extent to which source and target level sparse features overlap for mapped word pairs based on the 161-element subset of the Swadesh list. Setting λ to $0.1, 0.3$ and 0.5 resulted in approximately $20, 5$ and 2 non-zero coefficients per word on average. Performances in the 'No mapping' columns are extremely low. This is not surprising, however, since when

[1] https://sites.google.com/site/rmyeid/projects/polyglot
[2] https://hu.wikipedia.org/wiki/Swadesh-lista

no mapping between the source and target language embeddings is performed, overlap between the sparse representation can happen only due to chance.

Table 1 further reveals that the overlap in the sparse representation of the semantically equivalent word forms substantially increase due to any kind of mapping between the languages. The largest improvement can be observed in the case when the mapping matrix M is constrained to be orthonormal and the input word embeddings are made unit length and centered prior to the decomposition phase. As illustrated by the results, the orthogonality constraint mostly help to improve the recall of the mapped sparse word representations. Interestingly, when input word embeddings are left intact (i.e. no preprocessing is performed over them), requiring M to be orthonormal can slightly hurt performance.

	No mapping performed	Unconstrained M	Orthonormal M
$\lambda = 0.1$	0.043/0.036/0.039	0.139/0.112/0.124	0.114/0.112/0.113
$\lambda = 0.3$	0.023/0.012/0.016	0.169/0.122/0.141	0.117/0.109/0.113
$\lambda = 0.5$	0.008/0.004/0.006	**0.179/0.135/0.154**	0.138/0.127/0.132

(a) No preprocessing step performed on the input embeddings.

	No mapping performed	Unconstrained M	Orthonormal M
$\lambda = 0.1$	0.023/0.024/0.024	0.170/0.117/0.139	0.098/0.137/0.114
$\lambda = 0.3$	0.001/0.001/0.001	0.345/0.118/0.176	0.167/0.208/0.185
$\lambda = 0.5$	0.000/0.000/0.000	0.600/0.009/0.018	**0.271/0.202/0.232**

(b) Input embeddings made unit long and centered as a preprocessing step.

Table 1: The amount of overlap between source and target non-zero coefficients expressed in terms of Precision/Recall/F-score values for the elements of the Swadesh word lists induced dictionary.

3.2 Evaluation related to POS tagging

It has been shown previously that sparse word representations derived from the coefficient matrix can be beneficially utilized in various sequence tagging tasks such as POS tagging [5]. That paper evaluated sparse representations in the monolingual setting, however, here we investigate their applicability in the cross-lingual regime.

For comparability with prior work, we used the very same settings for training the sequence labeler, i.e. we generated features for the word forms by simply relying on their sparse features induced by the feature function ϕ. The linear-chain CRF models [15] we use are also trained using the crfsuite library [16] without modifying any of its default hyperparameter settings. Finally, evaluation was performed on the v1.2 Universal Dependencies treebank in terms of accuracy over the 17 element coarse-grained POS tag inventory it introduces.

The CRF models are trained on the English training data and evaluated on the Hungarian test set of the Universal Dependencies dataset without altering

any of the learned weights of the English model. The various CRF models trained on English data alone and evaluated on the English test have accuracies slightly below 0.9.

When judging the quality of applying the cross-lingually comparable sparse word representation, we can compare the evaluation scores achieved on the Hungarian test set to this mono-lingual setting. We should add, however, that performance drop during testing on the Hungarian test set not only originates from the insufficiencies of the cross-lingual representation, but also from the fact that we utilize the state transition features that are optimized to fit English texts. By defining the cross-lingual sparse representations, we are only able to adapt our test data in terms of the state features to the trained model.

Table 2 illustrates that applying the orthonormality constraint for the mapping matrix M improves POS tagging accuracies on the target language independent of the preprocessing step. This is a difference compared to the results of Section 3.1 where the orthonormality constraint only helped for the case when input embeddings were preprocessed prior to performing their decomposition.

	No mapping performed	Unconstrained M	Orthonormal M
$\lambda = 0.1$	0.111	0.300	0.315
$\lambda = 0.3$	0.034	0.283	**0.395**
$\lambda = 0.5$	0.096	0.204	0.384

(a) No preprocessing step performed on the input embeddings.

	No mapping performed	Unconstrained M	Orthonormal M
$\lambda = 0.1$	0.169	0.191	0.295
$\lambda = 0.3$	0.097	0.048	0.428
$\lambda = 0.5$	0.138	0.011	**0.446**

(b) Input embeddings made unit long and centered as a preprocessing step.

Table 2: Cross-lingual POS tagging accuracy with the pseudo-dictionary used to learn mapping M between languages.

Since this time we are not evaluating the quality of the cross-lingual sparse embeddings relative to the Swadesh lists, it is now possible to use them as a replacement for the noisy pseudo-dictionary for the calculation of the mapping matrix M. For that reason, we repeated our experiments on cross-lingual POS tagging such that the dictionary to construct M gets determined based on the 161-element dictionary induced from the English and Hungarian Swadesh lists. The results of this experiment can be seen in Table 3.

The most important observation to take when comparing Table 2 and Table 3 is that there is no substantial difference in the results whether the noisy or the more reliable word alignment is used to determine the inter-lingual mapping matrix M provided that the orhtonormality constraint is enforced for M. In fact,

there is a slight decrease of performance noticeable for using the more reliable word alignments when preprocessing of the word embeddings is also performed.

	No mapping performed	Unconstrained M	Orthonormal M
$\lambda = 0.1$	0.111	0.413	0.414
$\lambda = 0.3$	0.034	0.327	**0.431**
$\lambda = 0.5$	0.096	0.322	0.430

(a) No preprocessing step performed on the input embeddings.

	No mapping performed	Unconstrained M	Orthonormal M
$\lambda = 0.1$	0.169	0.372	0.338
$\lambda = 0.3$	0.097	0.310	**0.407**
$\lambda = 0.5$	0.138	0.310	0.356

(b) Input embeddings made unit long and centered as a preprocessing step.

Table 3: Cross-lingual POS tagging accuracy with the Swadesh-lists induced dictionary used to learn mapping M between languages.

For comparative purposes, we trained such CRF-models that make use of dense word embeddings. These results are included in Table 4. Comparing POS accuracies of the models which rely on sparse and dense word representations we can conclude that models relying on sparse features are dominantly better (but at least comparable) to those models which rely on dense word representation. Furthermore, the relative gain observed for mapping dense word embeddings in any way is less pronounced compared to those in the case of sparse word representations.

	No mapping performed	Unconstrained M	Orthonormal M
Pseudo-dictionary	0.156	0.039	0.144
Swadesh-based dictionary	0.156	0.346	0.264

(a) No preprocessing step performed on the input embeddings.

	No mapping performed	Unconstrained M	Orthonormal M
Pseudo-dictionary	0.201	0.010	0.328
Swadesh-based dictionary	0.201	0.294	0.368

(b) Input embeddings made unit long and centered as a preprocessing step.

Table 4: Cross-lingual POS tagging accuracy when relying on dense word representations.

4 Related work

The comprehensive survey in [4] enumerates a variety of approaches to map multiple independently trained (dense) word embeddings into the same embedding space. The approaches in the survey differ significantly in the amount of supervision they assume to have access to. In our work, we aimed at the utilization of as minimal external resources as possible by not relying on any parallel (or comparable) text resources between languages.

To this end, we created a highly noisy pseudo-dictionary similar to [12]. It has been showed that mapping noisily aligned signals close to each other is often worth to be performed by applying an orthonormal mapping [12,9,17,10].

The kind of decomposition we employ in this work differs both from [5] and [6] as here we not only require D to be a member of the convex set of matrices comprising of unit norm column vectors, but also enforce the coefficients in α to be non-negative.

5 Conclusions

In this paper, we introduced our approach for determining sparse word representations which are comparable across languages. We have shown that these cross-lingual sparse word representations can provide a substantially overlapping representation for word pairs with similar meaning of different languages. Our experiments demonstrated that such well behaving sparse cross-lingual representations can be obtained in the absence of parallel data across the languages, i.e. applying an automatically derived noisy dictionary performed on par to the scenario when a more reliable (but smaller) parallel dictionary was used for mapping word representations to the same embedding space.

Acknowledgement

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

1. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR **abs/1301.3781** (2013)
2. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics (2014) 1532–1543
3. Dozat, T., Qi, P., Manning, C.D.: Stanford's graph-based neural dependency parser at the conll 2017 shared task. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics (2017) 20–30

4. Upadhyay, S., Faruqui, M., Dyer, C., Roth, D.: Cross-lingual models of word embeddings: An empirical comparison. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, Association for Computational Linguistics (2016) 1661–1670
5. Berend, G.: Sparse coding of neural word embeddings for multilingual sequence labeling. Transactions of the Association for Computational Linguistics **5** (2017) 247–261
6. Faruqui, M., Tsvetkov, Y., Yogatama, D., Dyer, C., Smith, N.A.: Sparse overcomplete word vector representations. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, Association for Computational Linguistics (2015) 1491–1500
7. Al-Rfou, R., Perozzi, B., Skiena, S.: Polyglot: Distributed word representations for multilingual nlp. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning, Sofia, Bulgaria, Association for Computational Linguistics (2013) 183–192
8. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association of Computational Linguistics **5** (2017) 135–146
9. Xing, C., Wang, D., Liu, C., Lin, Y.: Normalized word embedding and orthogonal transform for bilingual word translation. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, Association for Computational Linguistics (2015) 1006–1011
10. Artetxe, M., Labaka, G., Agirre, E.: Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2016) 2289–2294
11. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. CoRR **abs/1309.4168** (2013)
12. Smith, S.L., Turban, D.H.P., Hamblin, S., Hammerla, N.Y.: Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In: Proceedings of the 5th International Conference on Learning Representations (ICLR 2017). (2017)
13. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online dictionary learning for sparse coding. In: Proceedings of the 26th Annual International Conference on Machine Learning. ICML '09, New York, NY, USA, ACM (2009) 689–696
14. Swadesh, M.: The origin and diversification of language: Edited post mortem by Joel Sherzer. Aldine, Chicago (1971)
15. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. ICML '01, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2001) 282–289
16. Okazaki, N.: CRFsuite: a fast implementation of Conditional Random Fields (CRFs) (2007)
17. Artetxe, M., Labaka, G., Agirre, E.: Learning bilingual word embeddings with (almost) no bilingual data. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics (2017) 451–462