

PAPER • OPEN ACCESS

Word-to-word Machine Translation: Bilateral Similarity Retrieval for Mitigating Hubness

To cite this article: Mengting Luo *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **533** 012051

View the [article online](#) for updates and enhancements.

Word-to-word Machine Translation: Bilateral Similarity Retrieval for Mitigating Hubness

Mengting Luo^{1,2}, Linchao He^{1,2}, Mingyue Guo^{1,2}, Fei Han^{1,2}, Long Tian^{1,2}, Haibo Pu^{1,2,a} and Dejun Zhang³

¹Lab of Agricultural Information Engineering, Sichuan Agricultural University, 0086-625014 Yaan, China

²Key Laboratory of Agricultural Information Engineering of Sichuan Province, 0086-625014 Yaan, China

³Faculty of Information Engineering, China University of Geosciences, 0086-430074 Wuhan, China

sookie0331@icloud.com

Abstract. Nearest neighbor search is playing a critical role in machine word translation, due to its ability to obtain the lingual labels of source word embeddings by searching k Nearest Neighbor (k NN) target embeddings from a shared bilingual semantic space. However, aligning two language distributions into a shared space usually requires amounts of target label, and k NN retrieval causes hubness problem in high-dimensions feature space. Although most the best- k retrievals get rid of hubs in the list of translation candidates to mitigate the hubness problem, it is flawed to eliminate hubs. Because hub also has a correct source word query corresponding to it and should not be crudely excluded. In this paper, we introduce an unsupervised machine word translation model based on Generative Adversarial Nets (GANs) with Bilingual Similarity retrieval, namely, Unsupervised-BSMWT. Our model addresses three main challenges: (1) reduce the dependence of parallel data with GANs in a fully unsupervised way. (2) Significantly decrease the training time of adversarial game. (3) Propose a novel Bilingual Similarity retrieval for mitigating hubness pollution regardless of whether it is a hub. Our model efficiently performs competitive results in 74min exceeding previous GANs-based models.

1. Introduction

Monolingual word distributed representation has great effect on majority Natural Language Process (NLP) applications, text classification [1], sentimental analysis [2], machine translation [3], short text similarity analysis [4] and so on. In addition, recent years witness an upsurge of research in bilingual embeddings learning which embedding multiple but independent monolingual word vectors in a shared semantic space. Bilingual embeddings learning is able to implement bilingual corpora semantic mining and similarity calculation. It extends the applications of monolingual embeddings to cross-semantic text similarity computation and multi-lingual word machine translation.

A common application of bilingual embeddings learning is the word-to-word machine translation. It usually implemented by two steps: the first step is linear transformation to align bilingual embeddings distribution; the second step is retrieving the k Nearest Neighbors (k NN) as source queries lingual



labels to build bilingual dictionary or predict existing dictionary. At present, there are still two problems: (i) the lack of parallel data. The parallel corpora and the labels of target are difficult to obtain in minority languages and dialects. Furthermore, existing dictionaries cannot fully contain vocabularies with gradually changing meaning in this constantly changing world, especially cyber language, therefore, it is necessary to extract potential information from the original monolingual data. (ii) The hubness problem. The k NN retrieval in unified feature space after linear transformation is seriously affected by pollution of ‘hubs’. Hub elements appears in the k NN list more than any other points but seem to be less meaningful and more disturbing.

In order to solve the above existing problems, we propose an unsupervised GANs-based machine word translation model with bilateral similarity, namely, Unsupervised-BSMWT, for mitigating hubness pollution in mapped semantic space. The goal of GANs is to learn bilingual embeddings mapping in an unsupervised manner. The generator products a transformation matrix W to gradually narrow the distance between the source vector space and the target vector space. The discriminator detects whether random inputs are from target space or from transformed source space. Furthermore, we first consider the Bilateral Similarity (BS), covering both bilingual and monolingual. For bilingualism, the BS considers the distance of source query to target candidate embeddings. For monolingualism, the BS considers the distance of source query to back-translation embeddings conversely. The BS can mitigate hubness problem effectively and make hubs become target candidates under the same probability instead of being rudely excluded. The structure of this paper is as follow: firstly, the relevant literature is briefly reported in section 2. Then, the proposed Unsupervised-BSMWT model is described in detail in section 3. We elaborate the overall word translation model in section 4 and discuss the experiment results in section 5. The paper closes with the main conclusions and future work in section 6.

2. Related work

With the focus on the unsupervised bilingual embeddings learning model for word-to-word machine translation, this section reports related research in several involving areas such as GANs for NLP [3], [5–7], the curse of dimensionality [8] and the best- k retrieval [9, 10].

2.1. GANs for NLP

Generative Adversarial Nets (GANs) proposed by Goodfellow et al. [11], is used to generate synthetic image. The GANs has been extensively exploited into a widely range of deep learning research. In speech recognition, GANs have capacity of generating ideal audio for data-hungry speech recognition system. In video representation learning, GANs is introduced to learn future-frame predictions replacing the conventional method of inherent pixel flows. Although the GANs shows competitive performance for computer vision, it is difficult to apply GANs for the field of NLP [12].

Recent years have witnessed a steady growth of the research in GANs for NLP. Li et al. [13] extends GANs to Markovian GANs for efficient real-time texture synthesis. Instead of generated natural language from a single corpus, Wang et al. [14] improve document embedding generation with multiple text corpora. In terms of Neural Machine Translation (NMT), Wu et al. [15] employ GANs to discriminate the results of NMT and human-translated sentences, which minimizes the gap between resulting sequence and expert translation. Yang et al. [16] improve NMT via conditional GANs, which could enable to regulate certain arbitrary external data. For the machine word translation, Zhang et al. [3] apply GANs for bilingual embeddings representation learning and alleviate the difficult in insufficient parallel data resources among minority language. Inspired by previous work, we apply GANs for machine word translation by transforming the source distribution to target in a unified semantic spacetext.

2.2. The curse of dimensionality

The k NN retrieval is an irreplaceable step to match transformed source query with appropriate translation in word-to-word machine translation. The best- k retrieval is affected by the intrinsic

property in high dimensionality, also called the curse of dimensionality [8]. In the past few years, the factor of dimensionality curse is studied thoroughly in early research (e.g. [17, 18]). For one thing, the Euclidean distance leads to a concentrated trend as the dimension increases. The phenomenon of concentrate shows that the Euclidean distance between all pairs of points is narrowed, making it difficult to measure the discrepancy. For another thing, the curse is relative to the k NN. The k NN [9], the simplest approach, selects target translations by ranking the best- k cosine similarity value of source query. The larger the value of cosine similarity between source query and target translation, the closer the meaning of the words is. But it causes many data points are ‘popular’ than other points [18] and may not even be qualitatively meaningful, resulting in the emergence of hubs [17]. Hubs have high occurrences rate in k NN list of all word vectors, which arises problems in a variety of high dimensional, data mining applications, in particular, indexing, information retrieval and classification clustering [17].

2.3. The best- k retrieval

To our best knowledge, in word-to-word translation, the best- translation candidate retrievals still relies on the Euclidean distance. The successful proposed approaches [7, 10, 19] all try to break the curse by one of the aspects that is relieving the hubness. The latest advanced approaches are as follows: Inverted Softmax (ISF) [10] applies softmax to normalize the probability of candidates and reverses search space to source space for detecting hubs and reduces their probability. Cross-domain Similarity Local Scaling (CSLS) [7], a bi-partite neighborhood graph approach, measures cosine similarity and the average cosine similarity of NN surrounding opposite language of source query and target candidates. That is to say, CSLS is to avoid candidates from dense areas that are heavily polluted by hubs. However, the apparent defect of above retrieval is getting rid of hubs from the list of candidates. Because hubs also have corresponding correct translations. In this paper, we mainly focus on the perspective of mitigating the hubness problem too. As will be described in section 4.3 we propose Bilateral Similarity retrieval form mitigating hubness pollution without excluding hubs.

3. Model overview

The Unsupervised-BSMWT is comprised of two main components in a fully unsupervised learning scheme: the GANs and the Bilateral Similarity (BS) retrieval. Firstly, the pre-trained source and target vectors are embedded into an identical feature space. Secondly, GANs generates a transformation matrix for bilingual embeddings learning. Finally, the BS retrieval extracts bilingual dictionaries from mapped embeddings.

Generative Adversarial Nets. On the first step, we introduce GANs for unsupervised bilingual distributions alignment. The GANs contains two sub-nets the generator nets and the discriminator nets. The work of generator is to produce a transformation matrix that allows mapped source embeddings gradually similar to target embeddings. While the work of discriminator is to detect random inputs which are from source or target semantic space. In every round of cat-chase-mouse competitive game, the generator runs once while the discriminator runs c times alternatively.

Bilateral Similarity retrieval. On the second step, we apply proposed BS retrieval for the best- k translation candidate retrievals in the mapped semantic space. The BS first select the k nearest neighbors of source query as target candidates, then the candidates are translated back as back-translation, finally BS measures more much appropriate translation candidates via combining distance of source query to target candidate and source query to back-translation.

Our contributions can be summarized as follows:

- 1) Innovation. We proposed a novel Bilateral Similarity for the best- k target candidate retrievals to alleviate hubness pollution. Particularly, BS measuring the probability of candidates via bilingual and monolingual distance that allows model hubs to not be excluded rudely. Moreover, BS is able to retrieve more appropriate target translation candidates especially in distance language

- 2) Efficiency. Our Unsupervised-BSMWT is **12.9×~19.4×** faster than previous GANs-based word translation model in training phase, which is computationally efficient.
- 3) High Performance. Our Unsupervised-BSMWT obtains competitive performance on various tasks in a fully unsupervised way. It reaches precision@1 79.87% translating English to Spanish without parallel sign and refinement.

4. Unsupervised-BSMWT

In this paper, we focus on developing GANs-based word machine translation with BS for bilingual distribution learning and word translation. The Unsupervised-BSMWT model is divided into three parts: (1) GANs for linear transformation by the generated transformation matrix W . (2) The nearest orthogonal restriction for W . (3) The Bilateral Similarity retrieval for bilingual lexicon induction.

4.1. GANs for linear transformation bilingual embeddings

We defined $X_s = \{x^{(i)}\}_{i=1,2,\dots,n}$ and $Y_t = \{y^{(j)}\}_{j=1,2,\dots,n}$ where $x^{(i)}$ and $y^{(j)}$ represent the lexical item from two languages. X_s and Y_t are two contacted matrices of every word embedding in their vocabularies. Where the subscript s and t represent the row of X and Y respectively come from source and target semantic space. On the base of early research [9], the similar structures of the distribution, representation of different language allows cross-lingual distribution learning via linear transformation

Generator layer: On the one hand, the goal of generator, $G(x_s : \theta_g)$, is to generate a transformation matrix W . The X_s is linearly transformed by W mapped into WX_s which is approximate to target semantic space Y_t .

Discriminator layer: On the other hand, the goal of discriminator, $D(y_t : \theta_d)$, is to find out lexical items that are not from target semantic space Y_t . Under the guidance of discriminator, the generator can product more deceptive W to narrow the distance between WX_s and Y_t .

Lost function: Our GANs contain two sub-nets the generator and discriminator networks. The objective function of the generator and discriminator is cross entropy loss, and two networks alternatively iterate with interval of c steps.

Lost function of discriminator as

$$\nabla \theta_d = -\frac{1}{m} \sum_{j=1}^m \log D(z(y_t^{(j)})) - \frac{1}{n} \sum_{i=1}^n \log D(1 - z(Wx_s^{(i)})). \quad (1)$$

Lost function of generator as

$$\nabla \theta_g = -\frac{1}{n} \sum_{i=1}^n \log D(z(Wx_s^{(i)})) - \frac{1}{m} \sum_{j=1}^m \log D(1 - z(y_t^{(j)})). \quad (2)$$

Where $z(\cdot)$ represents $N \sim (1, \sigma^2)$ Gaussian noise injection of random inputs [3]. The generated matrix W transform the source embeddings $x_s^{(i)}$ similar to $y_t^{(j)}$, and most source embeddings approximate to the corresponding target embeddings.

4.2. Orthogonal Constraint

After the adversarial game generating transformation matrix W , we recommend constraining W to orthogonal, for the following reasons.

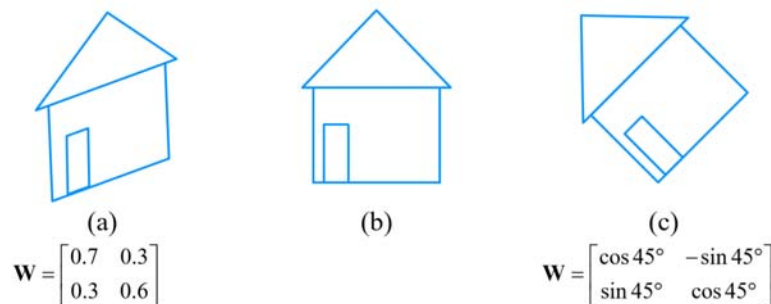


Figure 1. Bilateral Similarity retrieval. As small houses are depicted in this picture, we vividly treat all the edge of the house as word vectors and the house as matrix of contacting word embeddings. Both (a) and (c) are linear transformation forms of (b). (c) is the results of a orthogonal linear transformation, while (a) is non-linear.

4.2.1. Invariance

The orthogonal constraint preserves the invariance of matrix and the semantics of original data. As Fig.1 shown, the edge of little house can be thought of word embeddings. Like Fig.1(a), implemented with nonorthogonal linear transformation, the construction information of the Fig.1(a) is destroyed and the length and the angle of embeddings have changed larger or smaller. While Fig.1(c), implemented with orthogonal linear transformation, keeps the structure unchanged.

4.2.2. Self-consistent

An excellent translation model is required to have capacity of back-translation. Therefore, in order to translate back to, should be an identity matrix, and must be orthogonal. As shown in Fig.1(c), which is implemented with orthogonal linear transformation, can convert back to Fig.1(b) by generated, but Fig.(a) cannot be revert by.

4.2.3. Speed up training

Adopting orthogonal constraint effectively narrows down the search space, and therefore the constraint accelerates speed of convergence. Because the transformation matrix will never make the mapped source embeddings with non-orthogonal mapping weight like small Fig.1(a).

The calculation of orthogonal matrix is adopted [20] as

$$W^{r+1} = (1 + \beta)W^r - \beta(W^r W^{rT})W^r. \quad (3)$$

The updated is not strict orthogonal matrix, but an approximate orthogonal matrix. Where superscript r represents r -th state of in adversarial training, and we commend hyper-parameter for better performance.

4.3. Bilateral similarity for building dictionary

4.3.1. Target label retrieval in mapped space. As we discuss above, NN takes the nearest neighbors of mapped source, which causes the hubs contamination [21]. However, proposed approaches attempt to solve hubs problem by reducing the probability of hubs [10, 22] or by searching candidates in sparse space to avoid hubness contamination [7, 22]. Different from previous work, in order to address the challenge of relieving hubness pollution in high-dimensions and dense space without precluding target hubs, we proposed a novel Bilateral Similarity (BS) retrieval that considers both bilingual similarity and monolingual similarity.

Motivated by the game of passing on a message, BS treats the retrieval of searching the best- k translations as the messaging game among three agents 'A', 'B', 'A_B', as shown in Fig.2, the same icon represents same native language user. 'A' passes a message to 'B', and 'B' sends the message

back to 'A_B'. Then the two identical local language users ('A' and 'A_B') compare the information they received. The same game among 'A', 'C', 'A_C'. 'B' is the nearest neighbor in k NN list of 'A', but the back-translation 'A_B' is more similar with source query 'A'. While 'C' is inferior than 'B', but the back-translation 'A_C' is closer to 'A'. As Fig.2 illustrated, a nuance between 'A' and 'A_C' proves the candidate 'C' is highly reliable. On the contrary, farther distance between 'A' and 'A_B' proves that 'B' is highly disturbing. In particular, BS additionally considers the cosine similarity of source query to back-translation embeddings for the first time, which allows BS to distinguish between true interference candidates when hubs are not excluded.

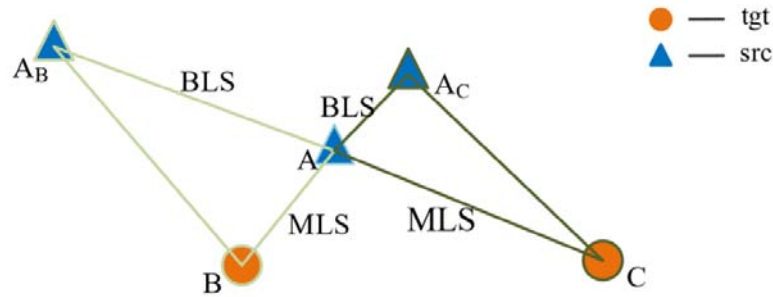


Figure 2. Bilateral Similarity retrieval. In the unified semantic space, the BS retrieves appropriate target candidates as semantic label of source query 'A'. 'A': source query. 'B': the top-1 NN. 'C': the top-2 NN. 'A_B': the backtranslation of 'A' from 'B'. 'A_C': the back-translation of 'A' from 'C'. BLS: BiLingual Similarity. MLS: MonoLingual Similarity. Suppose 'C' often appears in k NN lists of other points, regarded as a hub.

4.3.2. Formulation. The BiLingual Similarity (abbreviated as BLS to distinguish it from BS) is defined as the cosine similarity of source query embedding to target embedding as

$$\text{BLS}(Wx_s^{(i)}, y_t^{(j)}) = \cos(Wx_s^{(i)}, y_t^{(j)}) = \frac{Wx_s^{(i)} y_t^{(j)}}{\|Wx_s^{(i)}\| \|y_t^{(j)}\|}. \quad (4)$$

Similar to BLS, the MonoLingual Similarity (MLS) is defined as the cosine similarity of query source embedding $Wx_s^{(i)}$ to $Wx_{y_t^{(j)}}$, where $Wx_{y_t^{(j)}}$ is the back translation of $Wx_s^{(i)}$ from $y_t^{(j)}$:

$$\text{MLS}(Wx_s^{(i)}, Wx_{y_t^{(j)}}) = \cos(Wx_s^{(i)}, Wx_{y_t^{(j)}}) = \frac{Wx_s^{(i)} Wx_{y_t^{(j)}}}{\|Wx_s^{(i)}\| \|Wx_{y_t^{(j)}}\|}. \quad (5)$$

The weighted average of BLS and MLS between mapped source and target embeddings is defined as BS(...), which is represented as

$$\text{BS}(Wx_s^{(i)}, y_t^{(j)}) = (1 - \rho) \text{BLS}(Wx_s^{(i)}, y_t^{(j)}) + \rho \text{MLS}(Wx_s^{(i)}, Wx_{y_t^{(j)}}). \quad (6)$$

Where ρ is a hyper-parameter that is a continuous value between 0 and 1 balancing the two terms. After Bilateral Similarity retrieval in mapped space, the high-quality 1-to- k dictionaries are output without any given target label. The one-to-many dictionaries enhance the polysemy of semantic, and every target translation of source word vector $x_s^{(i)}$ can be calculated with identical transformation matrix W even if $x_s^{(i)}$ is not covered by resulting dictionary.

5. Experiment

We set two groups of experiments to show the performance of Unsupervised-BSMWT: (1) whether the Unsupervised-BSMWT learns cross lingual semantic without any parallel data and (2) evaluate the Unsupervised-BSMWT efficiency from both training time and translation accuracy. We report the results and discussion in each task.

5.1. Experiments Configuration

Word representations. Each monolingual embeddings of our model are trained using fastText [23] on the most 200,000 frequent words in Wikipedia with 300-dim, released by Facebook AI Research. We normalize the word embeddings to unit length before they are fed into GANs.

Training details. For generator, the size of W is obviously set for 300×300 to retain the dimension of embeddings after linear transformation. For discriminator, we construct two hidden layer of 2048 units with dropout rate of 0.1 and Leaky-Relu for non-linear activation. For discriminator regularization, we implement multiplicative Gaussian noise $N \sim (1, \sigma^2)$, where $\sigma = 0.5$. Then we alternately update the generator and discriminator using stochastic gradient descent with a batch size of 32 and the number of steps c is recommended $c = 5$, for hyper-parameters of BS, $\rho = 0.23$ usually works well.

5.2. Cross-lingual Semantic Similarity

The cross-lingual semantic similarity evaluates the relevance of the exported dictionaries by Unsupervised-BSMWT. High correlation proves that our model learns better bilingual embeddings. We apply the harmonic mean of Pearson correlation [24] for this task. The Pearson correlation refers to SemEval2017 [25] manually scored similarity scale by trained native or fluent speaker of target language user. And in this experiment, we refer NASARI [26], the official baseline of SemEval2017. As Tab.1 shown, our model achieves more than 0.7 Pearson correlation on all datasets without using parallel data. In comparison with NASARI, Unsupervised-BSMWT exceeds 0.11 score on English to German dataset, which learns more relevant bilingual semantic embeddings.

Table 1. Average pairwise Pearson correlation results of three datasets (EN to ES, DE and IT).

	EN-ES	EN-DE	EN-IT
NASARI [26]	0.64	0.60	0.65
Our	0.71	0.71	0.71

5.3. Bilingual lexicon induction

Bilingual Lexicon Induction is the most common application of word-to-word translation, and we report the training time and translate precision @k of our model on English to Spanish test set. In this task, the experiment is designed for testing the efficiency of speed and accurate, compared to previous GANs-based machine translate model. All experiments run on a single CPU.

Table 2. Word translation retrieval precision@1 for k=1. The results with symbol $\square$ use tiny data.

	EN-ES		ES-EN	
	time (min)	Precision @1	time (min)	precision @1
Zhang et al. (2017) [3]	1425	00.00	1425	00.00
Zhang et al. $\square$ (2017) [3]	78	68.76	78	72.78
Lamlpe et al. (2018) [7]	906	81.70	906	83.30
Our	74	79.87	74	79.20

It is noticed in Tab.2 that Zhang et al. [3] costs 1425min to learn the cross-lingual distribution embeddings of English to Spanish, but it does not seem to work at all. This method constructs a single hidden layer of discriminator with 500 size, fed tiny-scale datasets (5k most frequent word embeddings with 50-dim). While the discriminator of our network is set as two hidden layers with 2048 size, fed large-scale datasets (200k large-scope embeddings with 300-dim). We infer that the poor results are caused by small amounts of neural units that lacks the ability to fit large-scale datasets. As expected, Zhang's model works in tiny-scale datasets and obtains precision@1 68.76% costing 78min. The results are symbol with $\square$ in the Tab.2 taking almost equal time, our model achieves 8.77%

higher precision. In comparison with Lample's method, the precision of it is averagely exceeds ours 2.96%, but it costs $12.9\times$ time than our model. Overall, our model achieves competitive performance spending minimal time

6. Conclusion and future work

This paper develops a model of unsupervised bilingual embeddings learning for word-to-word translation including the transformation of source embeddings and the retrieval of source query. For the distribution transformation, we apply GANs to generate a transformation matrix in a fully unsupervised form. Then the matrix is restricted to be orthogonal for three factors, invariance, self-consistent and speed. Finally, we propose a novel Bilateral Similarity retrieval for searching the best-transformation. For the hubness problem, considering the bilateral similarity of bilingualism and monolingualism allows BS to detect real interference and select more appropriate candidates without excluding hubs. The results of experiments show that Unsupervised-BSMWT achieve competitive performance and significantly $19.4\times$ faster with the high efficiency, compared with previous GANs-based model. For future work, we will focus on further efficiency improvement and various refinement approaches in word translation.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant 61702350, and the Sichuan Province Department of Education (NO.17ZA0297).

References

- [1] X. Guo, H. Sun, T. Zhou, L. Wang, Z. Qu and J. Zang, *SAW Classification Algorithm for Chinese Text Classification*, 7, 2338–2352 (2015).
- [2] A. L. Maas, R. E. Daly, P.T. Pham, D. Huang, A.Y. Ng and C. Potts, *Learning Word Vectors for Sentiment Analysis*, 142–150 (2011).
- [3] M. Zhang, Y. Liu, H. Luan and M. Sun, *Adversarial Training for Unsupervised Bilingual Lexicon Induction*, 1, 1959–1970 (2017).
- [4] H. Zhang, G. Zhong, *Improving short text classification by learning vector representations of both words and hidden topics*, Knowl. Based Syst., 102, 76–86 (2016).
- [5] V. Premachandran, A. L. Yuille, *Unsupervised learning using generative adversarial training and clustering*, (2016).
- [6] A. Radford, L. Metz, S. Chintala, *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*, Int. Conf. Learn. Represent., (2016).
- [7] G. Lample, A. Conneau, M. Ranzato, L. Denoyer, H. Jegou, *Word translation without parallel data*, Int. Conf. Learn. Represent., (2018).
- [8] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag, (2016)
- [9] T. Mikolov, Q. V. Le, I. Sutskever, *Exploiting Similarities among Languages for Machine Translation*, Comput. Sci., (2013).
- [10] S. L. Smith, D.H.P. Turban, S. Hamblin, N.Y. Hammerla, *Offline bilingual word vectors, orthogonal transformations and the inverted softmax*, Int. Conf. Learn. Represent., (2017).
- [11] I. J. Goodfellow, J. Pougetabadie, M. Mirza, B. Xu, D. Wardefarley, S. Ozair, A. C. Courville, Y. Bengio, *Generative adversarial nets*, NLPs, 2672-2680 (2014).
- [12] L. Yu, W. Zhang, J. Wang, Y.Yu, *SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient*, AAAI, 2852-2858, (2017).
- [13] C. Li, W. Michael, *Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks*, Eur. Conf. Comput. Vis., 702–716 (2016).
- [14] B. Wang, D. Klabjan, *Generative Adversarial Nets for Multiple Text Corpora*, arXiv Comput, Lang, (2017).
- [15] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick. O.

- Vinyals, G. Corrado, M. Hughes, and J. Dean, *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*, abs/1609.0, (2016).
- [16] Z. Yang, W. Chen, F. Wang, B. Xu, *improving neural machine translation with conditional sequence generative adversarial nets*, north Am. chapter Assoc. Comput. Linguist., 1, 1346–1355 (2018).
- [17] M. Radovanovic, A. Nanopoulos, M. Ivanovic, *On the existence of obstinate results in vector space models*, 186–193 (2010).
- [18] M. Radovanovic, A. Nanopoulos, M. Ivanovic, *Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data*, J. Mach. Learn. Res., 11, 2487–2531 (2010).
- [19] G. Dinu, M. Baroni, *Improving zero-shot learning by mitigating the hubness problem*, arXiv Comput. Lang., (2014).
- [20] M. Cisse, P. Bojanowski, E. Grave, Y.N. Dauphin, N. Usunier, *Parseval Networks: Improving Robustness to Adversarial Examples*, Int. Conf. Mach. Learn., 854–863 (2017).
- [21] E. L. Denton, S. Chintala, A. Szlam, R. Fergus, *Deep generative image models using a Laplacian pyramid of adversarial networks*, neural Inf. Process. Syst., 1486–1494 (2015).
- [22] M. Artetxe, G. Labaka, E. Agirre, *Generalizing and Improving Bilingual Word Embedding Mappings with a Multi-Step Framework of Linear Transformations*, (2018).
- [23] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, *Enriching Word Vectors with Subword Information*, TACL, 5, 135–146 (2017).
- [24] D. Jurgens, M.T. Pilehvar, R. Navigli, *Semeval-2014 task 3: Cross-level semantic similarity*, In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), 17–26 (2014).
- [25] J. Camacho-Collados, M.T. Pilehvar, N. Collier, R. Navigli, *Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity*, In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (2017).
- [26] J. Camacho-Collados, M.T. Pilehvar, R. Navigli, *Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities*, Artif. Intell., 240, 36–64 (2016).