Correction of vector representations of words to improve the semantic similarity

Alexey Kolosov
Lomonosov Moscow State University
Moscow, Russian Federation
akolosov@cs.msu.ru

Archil Maysuradze
Lomonosov Moscow State University
Moscow, Russian Federation
maysuradze@cs.msu.ru

***Abstract* — The computational complexity of obtaining vector representations of words is constantly growing due to both an increase in the volume of input data and an increase in the complexity of the models. The design and training of a new model, which would take into account additional expert information about the semantic similarity between words, today is estimated at thousands of hours of computational experiments. In the study, existing vector representations are adjusted by transforming them into secondary representations taking into account the semantic similarity. The representation transformation and the learning procedure are proposed. Thus, instead of building a new model, the existing one is corrected, which significantly reduces computational costs compared to developing a new model.**

Keywords* — *vector representation, semantic similarity, distance realization

I. Introduction

In the subject area of Natural Language Processing, NLP, it is often necessary to represent segments of text with low-dimensional vectors that represent their semantics. If two words that are close in meaning can be represented by close vectors, then such representations can then be effectively used for a wide class of NLP problems, in particular, for the tasks of information retrieval, categorization and summarization of texts, sentiment analysis, determining the boundaries of named entities, resolving homonymy, generating responses in dialogue systems.

Historically first models of word representation did not involve learning. Today, in the overwhelming majority of cases, setting up a model involves optimizing some performance metrics on some training information. If training information of a new type appears, then it is necessary to change, if not the architecture of the model itself, then at least the perfomance metric and the method of its training. Thus, the full cycle of development and customization of a word representation model that takes into account an additional type of training information is very laborious.

In this paper, a compromise approach is considered, when, instead of creating a fundamentally new model of vector representation of words, the results of the existing model are corrected taking into account additional training information. As such additional information, expert estimates of the semantic similarity between words are used. The contribution of this work lies not only in the construction of a specific corrective operation, but also in the preliminary development of quality functionals and training methods for future richer models that will immediately use training information of different types. In addition, it is important to note that the presence of a correction operation is useful for objects that were not presented in the original sample.

Today widespread two types of models: context-free, which allow to obtain representations for isolated words [1], and context-depend, that give a representation of a word, taking into account a specific context [2]. Proposed approach is appropriated to both types of models: for models that allow obtaining context-depend representations of words, the performance metrics assumes the use of context-free representations of words. But since there is no labeled dataset for context-depend data, this type of representations was not included in the work.

The work is organized as follows: section 2 provides an overview of approaches to obtaining representations of words and sets of expert information; section 3 describes a method for improving the quality of vector representations of words using expert information and provides a justification of effectiveness of the proposed method; section 4 presents the implementation of the described approach, experimental study and it's results; section 5 summarizes the work.

II. Sources of semantic similarity

To correct the vector representations of words, two types of sources of information about the semantic similarity of words are needed. First, pre-computed vector representations of words are needed, obtained using one of the methods for automatic processing of the text corpus. Such representations are hereinafter referred to as primary. Secondly, expert datasets are used as additional information for adjusting primary representations, i.e. labeled by experts pairs of words with the value of their semantic similarity. Such sets are hereinafter referred to as expert similarity.

A. Primary representations

Context-independent representations of words used exactly one vector representation corresponds to each word, . Working with context-sensitive vector representations is a matter for further study. The following are common methods for obtaining this type of representations.

To construct vector representations of words with the property of semantic similarity, which means that similar in meaning words are compared vector representations close in the sense of a given distance function, the following approaches are used:

1) matrix decomposition

One of the problems with tf-idf approach is that the size of the word vector is equal to the size of the dictionary, which leads to computational difficulties when the size of the text corpus increases. The solution is the matrix decomposition of the word-document matrix into the product of two matrices: word-topic and topic-document. This approach is called topic modeling and in addition to basic decomposition methods such as SVD decomposition, more complex approaches have also been investigated that allow obtaining vector representations of words of comparable quality with the results of modern approaches [3]. An important advantage of thematic representations is the interpretability of individual components, due to the fact that each component can be associated with a specific topic.

Russian Foundation for Basic Research (project No. 20-01-00664-a)

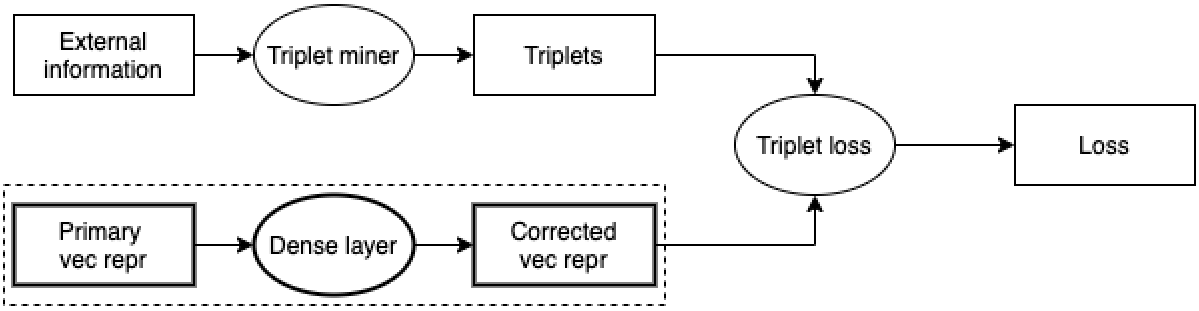

Fig 1. Solution pipeline.

2) word2vec

The word2vec approach is based on the locality hypothesis - “words that occur in the same environment have similar meanings”. The probability of a word is predicted from its context. As a result, such vectors of words are trained so that the probability assigned by the model to a word is close to the probability of encountering this word in this environment in a real text. This approach has shown its usefulness in practical problems of text analysis and in other subject areas. Its advantages include a relatively high quality indicator for well-known expert datasets. There is also a property of analogies, the work with which is a question for further research.

3) fasttext

Also worth noting is the approach to using the literal representation of a word to build word embedding [4]. The advantage of this approach is the ability to work with words that are not originally in the dictionary. However, it is not clear what can serve as external information in case of this approach.

B. *Expert essesments of similarity*

Expert information is specified as a set of word pairs and essesments in the format <word, word, essesment>. For each set, it is important to determine the number of pairs of words, the composition by parts of speech, the interval in which the essesment is put, the meaning of the essesments and the method of assessment — i.e. the guidelines given to the assessors. Distinguish two concepts: word similarity in the sense of proximity, i.e. synonyms, and word relatedness in the sense of domain proximity, i.e. words of the same topic.

1) WordSim353

The set was released in 2002 [5]. 353 word pairs in the dataset are divided into two parts: part of 153 pairs is marked by 14 experts, the rest of 200 words are marked by 16 experts. Each expert was asked to rate on a scale of 0 to 10 the relatedness of the words in a pair, and then the essesment was averaged among the experts. The original dataset contains the essesments of each specific expert, which can be used to obtain a better essesment than the usual averaging. For example, one can take into account the essesment of each expert with a weight that corrects the bias of the expert's essesments in one direction or another.

A significant disadvantage of this set is the use of only nouns. In addition, the instructions do not explain the difference between similarity and relatedness, so there is some confusion between these concepts, which results in insensitivity to this difference in dataset essesments.

2) MEN

The MEN Test Collection is a set with 3000 word pairs, collected in 2012 [6]. This set contains words of three parts of speech: nouns, verbs and adjectives, and words of different parts of speech can be found inside a pair. The essesments for pairs are obtained in a different way than in WordSim353: instead of asking the expert to essess the connectedness in a pair on a certain scale, the expert is asked to choose between two pairs, in which words are more connected. The assessment of each pair is obtained as a result of the work of one expert, there may be potential problems, there is no agreement of opinions of several experts. Each expert is given a pair of words and 50 randomly selected other pairs, and then 50 comparisons are expected from the expert in which pair of words are more related. As a result, in the case of the most related words in a pair, the expert will select this pair in comparison in all cases. Thus, the couple will receive the highest essesment of 50.

The key disadvantage of this dataset remains the same - similarity is considered by the authors to be a special case of relatedness. Experts again do not explain the difference between these concepts, so artifacts related to insensitivity to similarity and relatedness may appear in the set.

3) SimLex-999

This dataset was released in 2014 [7]. Initially, the set was conceived to solve the key problem of WordSim353 and MEN - insensitivity to the difference in similarity and relatedness. The principle of obtaining essesments from the opinions of experts here is the same as in WordSim353, but experts now clearly explain in the instructions the difference between similarity and relatedness, experts are asked to estimate exactly similarity. For example, for similar words ’coast’ - ’shore’ the essesment is 9.00 (SimLex-999) and 9.10 (WordSim353), and for related words ’clothes’ - ’closet’ the essesment is 1.96 (SimLex-999) and 8.00 (WordSim353).

This set contains words of different parts of speech: 666 pairs with nouns, 222 pairs with verbs, 111 pairs with adjectives. The disadvantage here will be the lack of pairs with different parts of speech in a pair, unlike MEN.

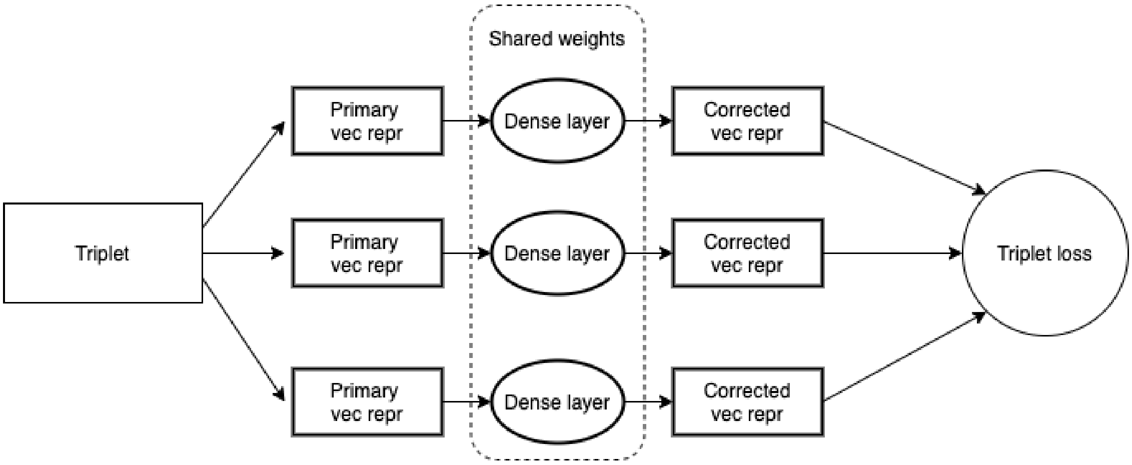

Fig 2. System step for triplet.

III. Correction of vector representations

Today, expert information on the semantic similarity of words is used only to assess the quality of already constructed vector representations of words, but not for the construction itself of such representations. In this paper, a certain compromise is considered: expert information about semantic similarity is used to correct vector representations that were originally built without it.

Expert information about the semantic similarity of words has a number of peculiarities. First, it is given on a relatively small number of words. Secondly, not all possible combinations of words form pairs. Thirdly, the scale of similarities is not arithmetic: the order of the assessments is meaningful, but not the absolute values. These peculiarities must be taken into account when developing the performance metrics and learning methods. These three peculiarities lead to the fact that one need to look not absolute values, but orders of values. So one need a performance metric that works with orders.

The proposed method is somewhat similar to the well-known triplet loss approach. The features consist in the formation of groups of objects and the calculation of the loss function on them. The quality functional is the average of all losses for all formed groups of objects. Conversion of primary vector representations into secondary ones is learned (with one dense layer). The form of this transformation allows for a balance between primary representations and additional requirements. The selection of groups of objects is determined by the sparseness of word pairs for which expert information is available. The loss function is based on the order of expert essesments, but not their absolute value, which corresponds to the tasks of realizing distances [8].

The scheme of the described approach is shown in Fig 1. Exactly the setting of the weights in the dense layer determines the corrected vector representations. However, from the point of view of the calculator, each corrected representation is determined not once, but many times, since despite the fact that the weights are shared by all representations, when calculating losses in triplets, it is necessary to redefine the adjusted representations each time. The scheme for calculating the loss for one triplet is shown in Fig 2.

The effectiveness of the approach is due to the fact that, just as when sorting a sequence of numbers by swap in incorrectly ordered pairs, each such permutation improves the overall sortedness of the entire sequence, also in the case of triples, changing the order to correct in one triple improves the overall order, since the quality indicator is considered as Spearman or Kendall rank correlation on sets of essesments for word pairs, which means it is necessary to sort the essesments. Each correction in a triple word results in an improvement in the overall ordering across the entire set of word pairs. Unfortunately, triplets alone are not enough to restore the exact order. To do this, one need to use quadruplets, the use of which is a question for further research.

The distance function is used to determine the closeness of the representetions. Often, some distance function is associated with the method of training representations, i.e. representations by construction are consistent with a certain function of distance. But not all studies interpret this distance as exactly semantic. It is also helpful to remember that most distance functions are subject to the curse of dimensionality. This has led to a significant predominance of cosine similarity for vector representations. But, generally speaking, several different distance functions can be used simultaneously on the same representations. As a result, it is important that there is agreement between the distance function used in the loss function (internal distance function) and the distance function that is applied to the resulting representations in the quality scoring (external distance function).

IV. Implementation and Experiments

The source of inspiration for the implementation of the described approach is the neural network dimensionality reduction tool Ivis [9]. This tool also uses the triplet loss function, but the different neural network model used, another source for generating the triplets, and the way triplets are generated is new.

The dimension of the word representation must be specified before correcting, i.e. for ready models, the dimension is fixed. In the approach, the input and output dimensions match, since the goal is not to lower the dimension, but to increase the performance metric, which is the rank correlation with the expert set of word pairs. In one case, the euclidean distance was chosen as the external and

Primary representetions	Expert assessments	Dist function	Spearman corr before correction	Spearman corr after correction	Kendall corr before correction	Kendall corr after correction
Word2vec lemmatize	WordSim	euclidean	0.70	0.75	0.51	0.56
		cosine	0.70	0.64	0.51	0.45
	MEN	euclidean	0.76	0.92	0.55	0.75
		cosine	0.76	0.73	0.55	0.52
	Simlex	euclidean	0.41	0.60	0.28	0.43
		cosine	0.41	0.39	0.28	0.27
Word2vec non lemmatize	WordSim	euclidean	0.71	0.76	0.52	0.56
		cosine	0.71	0.64	0.52	0.45
	MEN	euclidean	0.74	0.91	0.53	0.75
		cosine	0.74	0.71	0.53	0.50
	Simlex	euclidean	0.39	0.61	0.27	0.44
		cosine	0.39	0.38	0.27	0.26

Fig 3. Vector correction results.

internal function of the distance, and in the other case, the cosine distance.

The primary representations are those obtained using the word2vec approach. These representations are obtained from the corpus of Wikipedia texts and have a dimension of 300. In one set of primary representations, words are lemmatized, in another they are left in their original form. All primary representations are taken from the storage of vector representations [10]. The described expert datasets are used: WordSim353, MEN, SimLex-999.

The experiment is organized as follows:

1) For the selected primary representations and the selected expert dataset, a given external and internal distance function, the quality of the primary representations with the expert dataset is assessed by calculating the rank correlation.

2) On a model consisting of one fully connected layer with input and output dimensions of 300, the triplet loss function is calculated, the triplet of object identifiers for which comes from the triplet miner generator, which forms triplets from the expert dataset so that in one triplet the second word is closer to the first than the third, while the second and third words are selected from the set of words, with which the first word is presented in the expert dataset.

3) After calculating loss on the secondary vector representations, the weights are adjusted and the process is repeated until 100 epochs have passed or the weights stop changing.

4) The quality of the resulting secondary vector representations with an expert dataset is assessed by calculating the rank correlation and the change in the quality of the representations is determined.

The experimental results are presented in Fig 3. Euclidean distance function showed increase in quality both for Spearman or Kendall rank correlation in all experiments. Cosine distance showed decrease also in all experiments. Unfortunately, when dividing expert datasets into train and test, the quality of secondary representations tends to 0. This means that in the course of changing the representations of those words that are presented in external information, the representations of the remaining words were violated. This is most likely due to the fact that since the expert set is small, the transform function remembered the best representations for the words that make up this set. The described problem can be solved by using other primary representations as external information, the number of which is comparable to the number of used primary representations. Also since word2vec representetionsare normalized, they lie on a sphere where the cosine and euclidean are monotone, and therefore the orders are the same

V. Conclusion

The paper describes a new method for correcting vector representations of words to improve their semantic similarity, which became possible due to the use of expert data not only to score the quality, but also to correct vector representations. The use of vector representations with a higher rank correlation will increase the quality of solutions to the NLP problems. In addition, the proposed model can be used as an intermediate stage in more complex models using vector representations of objects.

Acknowledgment

This work is carried out with the financial support of the Russian Foundation for Basic Research (project No. 20-01-00664-a) and the state budget theme for research work No. 5.1.21 of the Lomonosov Moscow State University.

The research is carried out with the financial support of the Intellect Nonprofit Foundation for the Development of Science and Education.

References

[1] Mikolov T., Chen K., Corrado G., Dean J. / Efficient Estimation ofWord Representations in Vector Space // In Proceedings of Workshop at ICLR. —2013a.

[2] Devlin Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina / BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // 2018

[3] Potapenko A., Popov A., Vorontsov K. / Interpretable Probabilistic Embeddings: Bridging the Gap Between Topic Models and Neural Networks // AINL: Artificial Intelligence and Natural Language Conference. Vol. 789 2017. P. 167–180.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin and Tomas Mikolov / Enriching Word Vectors with Subword Information // 2016 Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606, 2016.

[5] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin / Placing Search in Context: The Concept Revisited // ACM Transactions on Information Systems, 20(1):116-131, January 2002 http://gabrilovich.com/resources/data/wordsim353/

[6] E. Bruni, N. K. Tran and M. Baroni / Multimodal Distributional Semantics // Journal of Artificial Intelligence Research 49: 1-47. https://staff.fnwi.uva.nl/e.bruni/MEN

[7] Felix Hill, Roi Reichart and Anna Korhonen / SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation // Computational Linguistics. 2015 https://fh295.github.io/simlex.html

[8] F. Chung, M. Garrett, R. Graham, D. Shallcross / Distance realization problems with applications to Internet tomography // Journal of Computer and System Sciences. 2001. Vol. 63. No 3. P. 432-448.

[9] Benjamin Szubert, Jennifer E. Cole, Claudia Monaco & Ignat Drozdov / Structure-Preserving Visualisation of High Dimensional Single-Cell Datasets // Scientific Reports, Vol. 9, No. 1, 2019, P. 8914.

[10] Fares, Murhaf; Kutuzov, Andrei; Oepen, Stephan & Velldal, Erik / Word vectors, reuse, and replicability: Towards a community repository of large-text resources // Proceedings of the 21st NoDaLiDa, 2017. http://vectors.nlpl.eu/repository/