Investigating parallelograms inside word embedding space using various analogy test sets in various languages

Rashel Fam　　Yves Lepage
早稲田大学大学院 情報生産システム研究科
fam.rashel@fuji.waseda.jp　　yves.lepage@waseda.jp

Abstract

The idea of using analogy to assess the quality of word embedding spaces implies the existence of parallelograms between the four terms of an analogy. We investigate the presence of analogy parallelograms in various word embedding spaces for various languages by relying on analogies contained in several analogy test sets. In this paper, we report a negative result: no parallelogram is found. We also discuss another possibility to approach the word as a small n-sphere instead of being a point inside the embedding space. Thus an analogy is formed as a parallelogram between four n-spheres.

1 Introduction

Previous works, like [1] and [2], claimed that there are linguistic regularities in word embedding spaces, and even sentence embeddings [3]. These regularities emerge as parallelograms on hyperplanes in the embedding space. Figure 1 presents an illustration of the claim where the four terms in an analogy make a parallelogram in the embedding space.

This claim was challenged by [4]. When the space is curved as in differential manifolds, the equality $\vec{D} = \vec{C}+\vec{B}-\vec{A}$ will not hold for the analogy $A : B :: C : D$. They proposed a parallelogramoid procedure using geodesic shooting and parallel transport to explain the analogical relation between words along curvatures in Riemannian manifolds.

In this paper, we perform an investigation of analogies that possibly exist in word embedding spaces. To investigate the existence of parallelograms in embedding spaces, we perform experiments in discovering the analogies contained in various analogy test sets. We explore embedding spaces and try to extract all analogies from these analogy test sets.

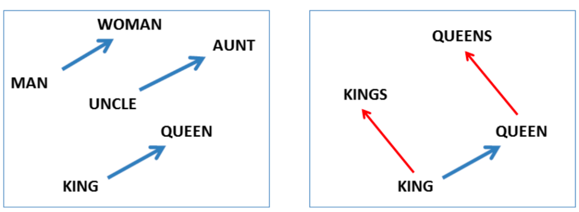

Figure 1 Parallelograms in word embedding space. Figure copied from [1]

2 Extraction of analogies from word embedding space

In the following section, we describe how the parallelogram defines an analogy and the implication of the definition on how we can extract analogies from a word embedding space.

2.1 Analogy as parallelogram

Figure 1 shows that linguistic regularity between four words, which is an analogy, makes a parallelogram. This parallelogram implies that there is an equality between ratios on the left hand and on the right hand of the analogy. For example, for the analogy *man* : *woman* :: *king* : *queen,* we should have the equality between the ratios of *man* : *woman* and *king* : *queen* (blue arrows in Fig. 1). In addition, to make a parallelogram, equality in the other direction is necessary: *man* : *king* = *woman* : *queen*. This is also true for analogy between other objects than words, like numbers. There is an equality be-

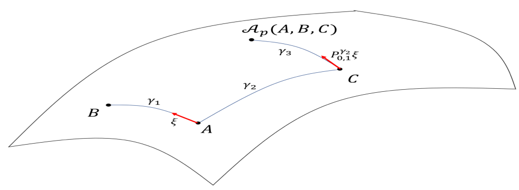

Figure 2 Parallelogramoid procedure on a Riemannian manifold. Figure copied from [4]

This work is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

tween the ratios $2:4$ and $3:6$ for the analogy $2:4::3:6$ (because of the properties of subtraction, the equality $2:3$ and $4:6$ is implied).

Unfortunately, previous works, like [1, 5], did not use this definition to solve the analogical equation: coin the word D given the tree words, A, B and C. After calculating the vector $\vec{D}$ from $\vec{A}$, $\vec{B}$ and $\vec{C}$, they will take vector $\vec{D'}$ which is the closest vector to $\vec{D}$. This is a relaxation of the claim that an analogy is a parallelogram. Prior to this, our intuition is that it will be very hard to have a true parallelogram inside the word embedding space.

Table 1 shows the result of a preliminary experiment on solving analogical equation of $man : woman :: king : x$ using three different pre-trained embedding models: fastText, word2vec and SENNA. This experiment was done in various languages: Belarussian (bel), Chinese (zho), French (fra), German (deu), Indonesian (ind), Javanese (jav), Sundanese (sun) and Thai (tha). The answers are checked by native speakers of the language. Results show that all of the answers are incorrect, except for French (fastText and word2vec), German (fastText) and Thai (fastText and word2vec).

Lang.	**Pre-trained embedding model**		
	fastText	word2vec	SENNA
bel	*звер-жанчына*	-	-
zho	万凰		
fra	*reine*†	*reine*†	*princesse*
deu	*königin*†	*Sibylle*	
ind	*kerajaan*	*rajanya*	-
jav	*Kirata*	*raja-raja*	-
sun	*Warmadewa*	-	-
tha	*ราชินี* †	*ราชินี* †	

Table 1 Solution of analogical equation $man : woman :: king : x$ in various languages using various pre-trained embedding models. A dagger mark (†) shows a correct answer according to human judgement (5 times out of 14). A hyphen ('-') means that there was no available pre-trained model for that particular language at the time the experiments were conducted.

2.2 Notions on analogy and extraction of analogical clusters

An analogy between four words, A, B, C and D, is noted as $A : B :: C : D$. The condition for an analogy to hold is an equality between the ratios. as shown in Formula (1).

$$A : B :: C : D \overset{\Delta}{\iff} \begin{cases} A : B = C : D \\ A : C = B : D \end{cases} \quad (1)$$

The ratio between two words, A and B, is defined as the difference of the vector representations of the words: $A : B \triangleq \vec{A} - \vec{B}$. We thus replace Formula 1 by Formula 2. With the difference between vectors, similarly as with numbers, the two equalities in the right part of Formula (2) are equivalent.

$$A : B :: C : D \overset{\Delta}{\iff} \begin{cases} \vec{A} - \vec{B} = \vec{C} - \vec{D} \\ \vec{A} - \vec{C} = \vec{B} - \vec{D} \end{cases} \quad (2)$$

Based on that, an analogical cluster is defined as a group of word pairs with the same ratio. This is basically the same as categories found in analogy test sets like capital-common-countries, currency, etc. (see Section 3).

$$\begin{matrix} A_1 : B_1 \\ A_2 : B_2 \\ \vdots \\ A_n : B_n \end{matrix} \overset{\Delta}{\iff} \begin{matrix} \forall (i, j) \in \{1, \dots, n\}^2, \\ A_i : B_i :: A_j : B_j \end{matrix} \quad (3)$$

3 Data

There are two main resources used in this work: pre-trained word embedding models and analogy test sets. We investigate whether the analogies contained in analogy test sets emerge as true parallelograms in pre-trained word embedding spaces.

3.1 Pre-trained word embedding models

We use fastText [6] pre-trained models. They provide models in various languages which allow us to compare across different languages. The models were trained on Common Crawl and Wikipedia using CBOW with position-weights. The models we used were trained in 300 dimensions, with character n-grams of length 5, a window of size of 5 and 10 negative samples.

Let us now turn to the distribution of values inside the vector. We sample 1,000 vectors from the embedding model. For each dimension of the vector, we calculate the mean and standard deviation. Figure 3 plots the means of values for the 300 dimensions of the fastText pre-trained model for English. We observe that the graph roughly follows a Gaussian distribution centered around zero. The error bars in the figure show the standard deviation of the mean. These error bars are not visible in the figure because most of the standard deviations are around 0.05, which is very small.

This work is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

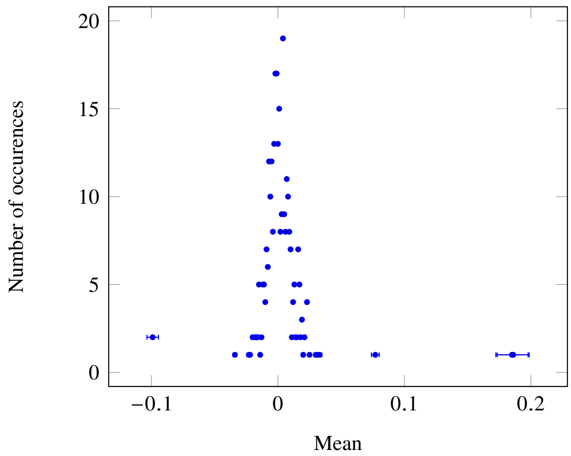

Figure 3 Distribution of fastText's vector means for each dimension with their standard deviation

3.2 Analogy test sets

We survey several analogy test sets that are publicly available. Table 2 shows the language availability of different analogy test sets.

- **Google** analogy test set[1)] [7] is probably the first analogy test set widely used since the emerging popularity of word embedding models. It contains general knowledge questions, like country-capital, and morphological questions, like singular-plural form of nouns. The analogy test set is originally only available in English.
- **fastText** analogy test set[2)] [8] is provided alongside the pre-trained models. The test set follows the format of its predecessor, Google analogy test set, and is available in French, Hindi and Polish.
- Bigger Analogy Test Set or usually called as **BATS**[3)] [9] is a bigger and more balance analogy test set in comparison to Google and fastText analogy test set. The analogy test set is also available for Japanese with the version called jBATS[4)] [10].
- Multilingual Generation of Analogy Datasets (**MGAD**)[5)] [11] is an analogy test set extracted from Universal Dependency treebanks. Thus, the analogical questions are restricted only to morphological phenomena. It is available in Hindi, Russian and Arabic.

Test set	**Language**						
	en	fr	hi	pl	ru	ar	ja
Google	✓						
fastText		✓	✓	✓			
BATS	✓						✓
MGAD			✓		✓	✓	

Table 2 Survey on the availability of analogy test sets

4 Experimental protocol

The purpose of our experiment is to investigate the existence of parallelograms inside the embedding spaces. We rely on analogy test sets as our ground truth. We investigate whether analogies contained in the analogy test sets actually make parallelograms. As the analogy test sets are already organised into categories, we check whether ratios in analogies that belong to the same categories are actually the same, i.e, whether one category makes one analogical cluster.

We carry out experiments in extracting analogical clusters from sets of words contained in each category of an analogy test set. Words are represented as vectors given by a pre-trained word embedding model. The extracted analogical clusters are expected to be similar to the categories contained in the analogy test set. To extract the analogical clusters, we use two different approaches.

The first approach relies on the strict definition of analogies where the equality of ratios has to hold in order to have an analogy. The algorithm to extract analogies from a given set of words is already presented elsewhere, such as [12, 13]. However, to ensure the equality of ratios, these techniques apply only to natural numbers (integer values). We convert the real values found on the vector dimensions into integer values by approximation, up to a certain precision after the decimal point. Formula (4) illustrates the approximation on a vector, with a precision of 3.

$$\begin{pmatrix} 0.1435 \\ 0.3496 \\ \vdots \\ 0.1180 \end{pmatrix} \Longrightarrow \begin{pmatrix} 143 \\ 349 \\ \vdots \\ 118 \end{pmatrix} \tag{4}$$

The second approach involves a common clustering algorithm. We perform DBSCAN clustering algorithm to cluster ratios. The reason behind it is the scalability and

1) http://download.tensorflow.org/data/questions-words.txt
2) https://fasttext.cc/docs/en/crawl-vectors.html
3) https://vecto.space/projects/BATS/
4) https://vecto.space/projects/jBATS/
5) https://github.com/rutrastone/MGAD

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

the geometry used (distances between points) which is aligned with the constraint that we use here with analogy. In this work, we use the implementation provided by scikit-learn [6] library.

5 Results and analysis

The results give no parallelogram found between words in the analogical test sets, as represented by vectors in any of the pre-trained embedding spaces considered, was found. This observation, which constitutes a negative result, gives support to the construction proposed in [4]. We also achieved the same result by using DBSCAN clustering algorithm.

5.1 A word as an area in the space

The analogy test sets are mainly used to assess the quality of a word embedding space. The test sets demand the embedding space to follow certain linguistic regularities, which are claimed to be semantical. However, in practice, some heuristics and tricks are introduced while performing the analogy task. For example, deleting the words included in the problem itself (the term A, B and C) from the candidates of the solution. Word D is enforced to be different than words A, B and C even when the true vector D that is calculated by the algorithm is closer to any of these words.

We propose that we should think of a word not as a point, but rather as a small n-sphere in the embedding space. By adopting this approach, we may find that this small n-sphere for a word may includes several words. The visibility and representation of the meaning of a word in the embedding space is extended by the proximity of the words in the neighbourhood. Thus, the analogy is now formed by the four small n-spheres instead of just four points in the embedding space. Here, we can imagine that the words *king*, *duke*, *prince*, *count*, etc. may have their extended n-sphere intersect or even inside each other. This makes the heuristics and tricks that we have done before sound natural.

5.2 Hypernymy and hyponymy

Capitalising on the approach of a word as a small n-sphere, here, we may get another explanation of how the embedding space outputs the solution which is the hypernymy or hyponymy of the true answer. For example, we may get one of the king's names instead of the word *king* itself. This varies depending on the corpus on which the embedding space is trained on. The discussion comes to whether there is any feature for the degree of generality of a word in embedding spaces; whether distributional semantics captures hyponymy and hyponymy. [14, 15] provides experiments on several datasets to observe whether hypernymy structures exist and are preserved inside the embedding space.

5.3 Task of analogy

Let us now reflect back on the task of analogy. It is important for us to ask ourselves again what are better analogies to design. One possible approach to answer that is to extract all possible analogies from a word embedding space. We need to have a critical view or be able to analyse these extracted analogies to draw conclusions about their validity of acceptability. Of course, we have to be more precise about the task at hand. If the goal is to assess the quality of the embedding space, then it is strictly demanded that the previously mentioned tricks are not fair.

6 Conclusion

By relying on analogies contained in analogy test sets, we investigated the existence of parallelograms inside a word embedding space model, fastText. The experiment consists in rediscovering the analogies by extracting analogies defined as the equality of ratios between the four terms. This implies that we only want true parallelograms. Experimental results showed that no analogy can be extracted from the word embedding spaces. We then applied a common clustering algorithm, DBSCAN, to extract the analogies. This way we allow for a loose parallelogram. This result supports the construction proposed in [4] where parallelograms for analogies are claimed not to exist in differential manifolds. Instead, they propose that analogies should follow the *Ricci* curvature rather than making parallelograms. In this paper, we discussed another way to approach the representation of a word in the embedding space: a word is not a point but a small area (n-sphere).

6) https://scikit-learn.org/stable/modules/clustering.html

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

Acknowledgement

The work reported here was supported by a JSPS Grant, Number 21K12038 (Kakenhi C), entitled ”Theoretically founded algorithms for the automatic production of analogy tests in NLP”.

References

[1] Tomas Mikolov, Wen-Tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT–2013)*, pp. 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[2] Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pp. 171–180, Ann Arbor, Michigan, June 2014. Association for Computational Linguistics.

[3] Xunjie Zhu and Gerard de Melo. Sentence analogies: Linguistic regularities in sentence embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3389–3400, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[4] Pierre-Alexandre Murena, Antoine Cornuéjols, and Jean-Louis Dessalles. Opening the parallelogram: Considerations on non-euclidean analogies. In *Proceedings of the 26th International Conference (ICCBR–2018)*, pp. 597–611, Stockholm, Sweden, 07 2018.

[5] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[6] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[7] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[8] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[9] Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn’t. In *Proceedings of the NAACL-HLT SRW*, pp. 47–54, San Diego, California, June 12-17, 2016, 2016. ACL.

[10] Marzena Karpinska, Bofang Li, Anna Rogers, and Aleksandr Drozd. Subcharacter Information in Japanese Embeddings: When Is It Worth It? In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pp. 28–37, Melbourne, Australia, 2018. Association for Computational Linguistics.

[11] Mostafa Abdou, Artur Kulmizev, and Vinit Ravishankar. MGAD: Multilingual generation of analogy datasets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[12] Yves Lepage. Analogies between binary images: Application to Chinese characters. In Henri Prade and Gilles Richard, editors, *Computational Approaches to Analogical Reasoning: Current Trends*, pp. 25–57. Springer, Berlin, Heidelberg, 2014.

[13] Rashel Fam and Yves Lepage. A study of the saturation of analogical grids agnostically extracted from texts. In *Proceedings of the Computational Analogy Workshop at the 25th International Conference on Case-Based Reasoning (ICCBR-CA–2017)*, pp. 11–20, Trondheim, Norway, June 2017.

[14] Zheng Yu, Haixun Wang, Xuemin Lin, and Min Wang. Learning term embeddings for hypernymy identification. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI’15, p. 1390–1397. AAAI Press, 2015.

[15] Ivan Sanchez and Sebastian Riedel. How well can we predict hypernyms from word embeddings? a dataset-centric analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 401–407, Valencia, Spain, April 2017. Association for Computational Linguistics.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).