GENERATING SEMANTIC SIMILARITY ATLAS FOR NATURAL LANGUAGES

Lütfi Kerem Şenel[1,2,3], *İhsan Utlu*[1,2], *Veysel Yücesoy*[1], *Aykut Koç*[1], *Tolga Çukur*[2,3,4]

[1]ASELSAN Research Center, Ankara, Turkey
[2]Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey
[3]Sabuncu Brain Research Center, UMRAM, Bilkent University, Ankara, Turkey
[4]Neuroscience Program, Bilkent University, Ankara, Turkey

ABSTRACT

Cross-lingual studies attract a growing interest in natural language processing (NLP) research, and several studies showed that similar languages are more advantageous to work with than fundamentally different languages in transferring knowledge. Different similarity measures for the languages are proposed by researchers from different domains. However, a similarity measure focusing on semantic structures of languages can be useful for selecting pairs or groups of languages to work with, especially for the tasks requiring semantic knowledge such as sentiment analysis or word sense disambiguation. For this purpose, in this work, we leverage a recently proposed word embedding based method to generate a language similarity atlas for 76 different languages around the world. This atlas can help researchers select similar language pairs or groups in cross-lingual applications. Our findings suggest that semantic similarity between two languages is strongly correlated with the geographic proximity of the countries in which they are used.

***Index Terms*—** cross-lingual semantic similarity; natural language processing; semantic similarity; word embedding, computational linguistics

1. INTRODUCTION

There are more than 7,000 languages spoken throughout the world, however, only 23 of them are used by more than half of the entire world population [1]. Most of the attention in NLP research is focused on this small portion of languages that are prevalent. With the increase in the use of data-driven methods, languages that lack sufficient resources have become more difficult to process. Especially after the rise of deep learning and neural network based methods that fundamentally require data, need for rich resources has become more evident. During the earlier years of the data-driven approaches, several studies [2–4] showed that NLP tools for low-resource languages can be improved by using the resources from resource-rich languages such as English. With the increasing popularity of neural models that generate monolingual word representations [5–8], called *word embeddings*, significant research effort is focused on learning *cross-lingual embedding models* in order to transfer knowledge from a resource-rich language to a low-resource language, and to represent meaning in cross-lingual applications. Among the studies that aim to learn cross-lingual embedding models, some try to learn transformation matrices to map monolingual word representations in one language to representations in another language using seed bilingual lexicons [9–12], while others learn cross-lingual embeddings directly from pseudo-aligned multilingual corpora [13, 14].

Many cross-lingual studies that work with low-resource languages [3, 15] show that working with similar languages provide improved performance compared to languages that are fundamentally different. Yet, this statement raises the question of how similarity between languages should be defined. Natural languages have complex structures that contain many different features such as phonology, morphology, word order and lexicon, upon which linguistic similarity may be defined. Different linguistic features might have different levels of influence on performance for different NLP tasks. For instance, while phonological features are significant for speech recognition, syntactic features and word order can be more important for machine translation. In the literature, different features have been used to define the similarity between languages. Linguists use genetic relationships between languages [16] to define language similarity and to construct language family trees. Lexical similarity, which measures similarity in both form and meaning [1], is used to measure the mutual intelligibility between different languages. In another study [15], typological features extracted from World Atlas of Language Structures [17] are used to define a similarity metric and to induce language clusters.

Many NLP applications such as sentiment analysis [18],

Lütfi Kerem Şenel: lksenel@aselsan.com.tr
İhsan Utlu: utlu@ee.bilkent.edu.tr
Veysel Yücesoy: vyucesoy@aselsan.com.tr
Aykut Koç: aykutkoc@aselsan.com.tr
Tolga Çukur: cukur@ee.bilkent.edu.tr

T. Çukur and A. Koç mutually supervised this work under a joint industry-university co-advising program.

word sense disambiguation [19] and measuring text similarity [20], significantly rely on semantic information and make use of representation spaces that accurately reflect semantic relations between words. Therefore, on such semantics-driven tasks, a low-resource language may benefit notably more from the incorporation of a resource-rich language that is particularly similar to it in terms of its semantic features compared to the contrary. In order to find the optimal language pairs or groups to work with, a similarity measure that focuses specifically on semantic features of the languages holds greater promise than the measures discussed above. In a recent study [21], representational similarity analysis (RSA) originally proposed to relate neural activity to computational models [22], is used to quantify the semantic similarities between 5 European languages. RSA computes the geometric similarity between two representation spaces by calculating the correlations between dissimilarity matrices, where dissimilarity matrix for each representation space is constructed by calculating the distances between samples (words in this case) using some distance metric. Since word embedding spaces are shown to represent semantic relationships between words accurately, RSA can be used on word embeddings to measure semantic similarities between languages.

This paper aims to construct a semantic similarity atlas for 76 different languages across the world that can be used to select language pairs and groups for cross-lingual studies. Semantic similarities are calculated using RSA on pre-trained fastText word vectors [8] based on a seed lexicon of 2443 English words that are translated to other languages using Google Translate.

This paper is organized as follows: In Section 2, we describe the methods used to construct the similarity atlas. Then, we present our findings in Section 3 and conclude this paper in Section 4.

2. METHODS

2.1. Word Vectors and Lexicon

Wikipedia is commonly used in multilingual and cross-lingual NLP studies due to its multilingual characteristics [23, 24]. Since this study focuses on measuring semantic similarity between languages, having a compatible source corpus is critical for the reliability of the results. There are three popular monolingual word embedding algorithms, word2vec [5], GloVe [7] and fastText [8] that are commonly used to represent meaning of words in a continuous space where the meaning is encoded by the relative positions of words with respect to other words in the vocabulary of a language. Among these three algorithms, fastText is the latest and it claims to generate state-of-the-art word representations especially for the morphologically rich languages due to subword information it utilizes. Moreover, the authors provide pre-trained word vectors that are trained on Wikipedia articles for 294 different languages. Pre-trained fastText word vectors have, therefore, stood out as a good choice for the source representations for the languages.

A seed multilingual lexicon is required in order to apply RSA to word vectors and characterize semantic similarities and dissimilarities between languages. For this purpose we use the lexicon introduced in [21] due to its sufficiently large size (2443 words) and the broad topic coverage. This allows the proposed method to consider semantic relations between words from many domains. In order to translate source lexicon, which is in English, to other languages, we use Google Translate tool. However, Google Translate does not provide translations for most of these 294 languages. Moreover, most of these languages do not have sufficiently large Wikipedia content to learn word vectors of sufficient quality. Therefore, the scope of this study is limited to 76 languages that Google Translate provides translation service, and that have more than 10,000 Wikipedia articles at the time of access.

Some of the words in the seed lexicon do not have corresponding single word expressions or their translations are not included in the corresponding fastText vocabulary for some of the target languages. Since number of words in the seed lexicon that have corresponding word vectors in all other languages is nearly zero, different subsets of the original seed lexicon are used for different language pairs.

2.2. Representational Similarity Analysis (RSA)

Representational similarity analysis was introduced in [22] in order to quantitatively relate neural activity measurements to computational theory by comparing their computed representational dissimilarity matrices (RDMs). For a language, RDM is taken as the symmetric matrix consisting of the pairwise cosine distances between vectors corresponding to the words in the lexicon as described in [21]. RDM thus represents the semantic structure of a language in terms of pairwise word similarities expressed in the form of word vector distances in the embedding space. Therefore, languages with similar semantic properties are expected to have similar RDMs.

Lexicons of different languages have different sizes due to invalid translations (translations to multiple words or to out-of-vocabulary words). Therefore, in order to calculate the correlations between RDMs for different language pairs, first the subset of the original lexicon that only contains words that are common in lexicons of both languages in the language pair is determined. Then, semantic similarity between a language pair is taken as the correlation between the RDMs from the obtained common lexicon.

Same process is applied to all 2850 language pairs from 76 different languages and a resulting 76×76 symmetric semantic similarity matrix is obtained.

Table 1. Semantic similarities between 10 different languages[1]

	en	es	de	tr	az	zh	ar	ru	uk	kk
en	1	0.61	0.58	0.48	0.39	0.23	0.42	0.54	0.52	0.39
es	0.61	1	0.52	0.45	0.36	0.21	0.38	0.50	0.47	0.35
de	0.58	0.52	1	0.41	0.34	0.19	0.36	0.49	0.47	0.34
tr	0.48	0.45	0.41	1	0.43	0.13	0.34	0.42	0.41	0.37
az	0.39	0.36	0.34	0.43	1	0.06	0.29	0.36	0.38	0.40
zh	0.23	0.21	0.19	0.13	0.06	1	0.12	0.15	0.14	0.06
ar	0.42	0.38	0.36	0.34	0.29	0.12	1	0.35	0.33	0.30
ru	0.54	0.50	0.49	0.42	0.36	0.15	0.35	1	0.62	0.36
uk	0.52	0.47	0.47	0.41	0.38	0.14	0.33	0.62	1	0.38
kk	0.39	0.35	0.34	0.37	0.40	0.06	0.30	0.36	0.38	1

Table 2. Semantic similarity ranks of the 10 languages with respect to other languages

	en	es	de	tr	az	zh	ar	ru	uk	kk
en	-	1	6	27	45	70	40	11	17	49
es	4	-	8	26	45	70	41	9	22	47
de	1	3	-	31	47	71	43	10	15	45
tr	1	2	21	-	9	72	44	11	16	36
az	3	18	38	1	-	73	58	15	5	2
zh	4	9	15	32	57	-	37	21	26	59
ar	1	3	10	27	45	72	-	14	28	39
ru	3	8	11	31	40	71	43	-	1	39
uk	3	14	16	30	39	72	45	2	-	38
kk	4	27	34	9	3	73	53	14	6	-

3. RESULTS

Obtained semantic similarity matrix is too large to display and manually inspect. Instead, representative similarity matrix for 10 languages is displayed in Table 1. One point that can be noticed from Table 1 is that most of the languages have relatively high semantic similarities with English while their semantic similarities with Chinese are significantly lower. To investigate this result in detail, Table 2 is constructed. Each row in Table 2 lists the semantic similarity ranks of the 10 languages in the columns for the language corresponding to that row. From a different perspective, each column lists the semantic similarity rank of the language corresponding to that column for the languages in the rows. It can be clearly seen that English is among the top ranks of the other 9 languages presented, while Chinese is within the bottom ranks. Moreover, one should notice that the table is not even close to be symmetric. This signifies that there might be bias in the similarity measures towards or against some of the languages. It can be argued that the results are affected by the size of the initial corpora of the languages or by the quality of the translations. It is also possible that this result is not due to some imperfection in the measurement process but rather due to the inherit nature of the languages; that is, some languages may be inherently similar to many other languages whereas some of them may significantly differ from the others in terms of their semantic structure.

It is difficult to identify a specific reason for the above findings due to the large number of languages, resource and time limitations, and language barriers. Nevertheless, the resulting semantic similarity matrix can reliably be used to generate a two dimensional semantic similarity atlas of the 76 languages. The effect of the possible bias in the measurement is minimized when the rows (or columns) of the similarity matrix are considered as features for that language rather than individual similarity values. This is because, with this approach, for a language to be considered similar to another language, they must be close to each other in this feature space. In other words, they should carry comparable levels of similarities to other languages. Here, the degree of the similarity to be high or low is not the concern, but what is important is whether they have similar values or not. Now, we move on to obtain a language similarity atlas that can be used to select language pairs or groups to work with on cross-lingual NLP applications. To do this, 76 dimensional similarity space is reduced to two dimensions by using the t-distributed Stochastic Neighbor Embedding (t-SNE) method that is commonly used to visualize high-dimensional data.

Figure 1 displays the resulting 2-dimensional language semantic similarity atlas. First observation one can make from the atlas is that languages that are spoken in geographically closer countries show higher semantic similarity in general, making the atlas resemble a geographic atlas. Throughout the history of languages, stronger interactions observed among the neighboring countries. Therefore, it is reasonable to have higher semantic similarities between languages spoken in neighboring countries. Although presented 2-dimensional atlas provides promising results that can potentially help researchers to work on similar languages, original 76 dimensional feature space without dimensionality reduction can be used for more accurate comparisons between languages. We make the full semantic similarity matrix publicly available [2] along with the translated lexicons for each language for other researchers to use in their studies.

[1] en: English, es: Spanish, de: German, tr: Turkish, az: Azerbaijani, zh: Chinese, ar: Arabic, ru: Russian, uk: Ukrainian, kk: Kazakh

[2] github.com/avaapm/LanguageSimilarityAtlas

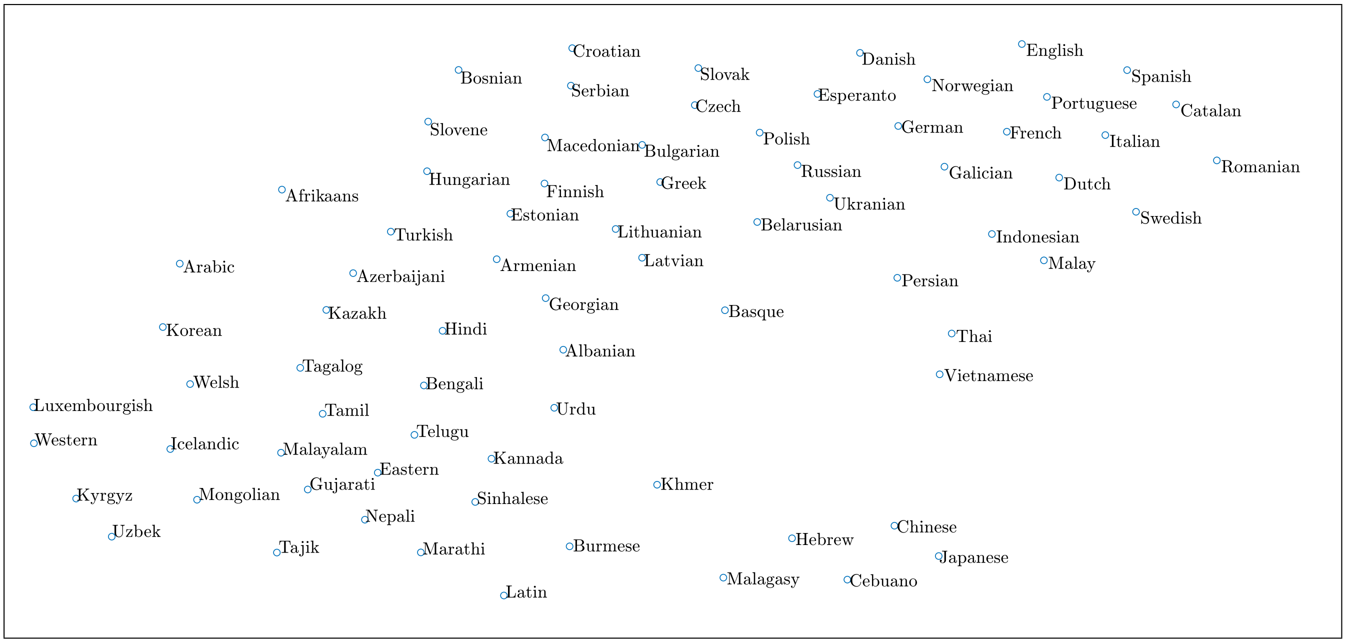

Fig. 1. Semantic similarity atlas for 76 different languages. 76 dimensional semantic similarity matrix obtained via RSA is reduced to two dimensions using t-SNE method. Languages that are spoken in geographically closer countries show higher semantic similarity in general. Therefore the atlas resembles a geographic atlas.

4. CONCLUSION

In this study, cross-lingual semantic similarities between 76 languages around the world are quantified using representational similarity analysis, and the resulting matrix is used to obtain a semantic similarity atlas. Pre-trained fastText word vectors trained on Wikipedia are used as source representations for the languages. Representational dissimilarity matrices (RDMs) are constructed for each language based on pairwise distances between word vectors corresponding to words from a word list that is translated to each language using Google Translate. Then, semantic similarity between languages are taken as the correlations between the RDMs. Rows of the resulting 76 dimensional semantic similarity matrix are taken as features to prevent possible bias due to measurement process and dimensionality is reduced to 2 using t-SNE in order to obtain semantic similarity atlas for the languages. From the resulting atlas, it is observed that languages that are spoken in neighbouring or geographically close countries are semantically similar in general.

Significant research effort is focused on cross-lingual NLP applications, and it is shown that working with similar languages provide performance improvements. However, how similarity between languages should be defined has been an open question. Although different similarity measures have been proposed by researchers from different fields, a similarity measure that focuses on semantic structures of languages can be useful in selecting language pairs or groups to work with especially for the tasks requiring semantic knowledge, including sentiment analysis and word sense disambiguation.

In this paper, pairwise cross-lingual semantic similarities between 76 different languages around the world are quantified. The obtained results show that some languages such as English share a relatively high degree of semantic similarity with most of the other languages while some other languages such as Chinese share relatively low semantic similarities with other languages. Specific reasons behind this possible bias towards and against some languages can be investigated in a future study.

5. REFERENCES

[1] Gary F. Simons and Charles D. (eds.) Fenning, *Ethnologue: Languages of the World*, Dallas, Texas: SIL International, 2018.

[2] David Yarowsky and Grace Ngai, “Inducing multilingual pos taggers and np brackets via robust projection across aligned corpora,” in *Proceedings of NAACL*, 2001.

[3] Jiri Hana, Anna Feldman, and Chris Brew, “A resource-light approach to russian morphology: Tagging russian using czech resources,” in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004.

[4] Fei Xia and William Lewis, "Multilingual structural projection across interlinear text," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 2007, pp. 452–459.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., pp. 3111–3119. Curran Associates, Inc., 2013.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

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

[8] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

[9] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever, "Exploiting similarities among languages for machine translation," *arXiv preprint arXiv:1309.4168*, 2013.

[10] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni, "Improving zero-shot learning by mitigating the hubness problem," *arXiv preprint arXiv:1412.6568*, 2014.

[11] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin, "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*, 2015, pp. 1006–1011.

[12] Manaal Faruqui and Chris Dyer, "Improving vector space word representations using multilingual correlation," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, pp. 462–471.

[13] Min Xiao and Yuhong Guo, "Distributed word representation learning for cross-lingual dependency parsing," in *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, 2014, pp. 119–129.

[14] Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith, "Massively multilingual word embeddings," *arXiv preprint arXiv:1602.01925*, 2016.

[15] Ryan Georgi, Fei Xia, and William Lewis, "Comparing language similarity across genetic and typologically-based groupings," in *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 385–393.

[16] Merritt Ruhlen, *On the origin of languages: studies in linguistic taxonomy*, Stanford University Press, 1994.

[17] Matthew S Dryer, David Gil, Bernard Comrie, Hagen Jung, Claudia Schmidt, et al., "The world atlas of language structures," 2005.

[18] Cicero dos Santos and Maira Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 69–78.

[19] Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli, "Embeddings for word sense disambiguation: An evaluation study.," in *ACL (1)*, 2016.

[20] Tom Kenter and Maarten De Rijke, "Short text similarity with word embeddings," in *Proceedings of the 24th ACM international on conference on information and knowledge management*. ACM, 2015, pp. 1411–1420.

[21] Lutfi Kerem Senel, Veysel Yücesoy, Aykut Koç, and Tolga Çukur, "Measuring cross-lingual semantic similarity across european languages," in *TSP*, 2017.

[22] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini, "Representational similarity analysis-connecting the branches of systems neuroscience," *Frontiers in systems neuroscience*, vol. 2, pp. 4, 2008.

[23] Lei Zhang, Achim Rettinger, and Steffen Thoma, "Bridging the gap between cross-lingual nlp and dbpedia by exploiting wikipedia," *NLP & DBpedia*, 2014.

[24] Alexander E Richman and Patrick Schone, "Mining wiki resources for multilingual named entity recognition," *Proceedings of ACL-08: HLT*, pp. 1–9, 2008.