Syntactic similarity of the sentences in a multi-lingual parallel corpus based on the Euclidean distance of their dependency trees

Masanori Oya
Meiji University
masanori_oya2019@meiji.ac.jp

Abstract

This study proposes the idea that the difference between the syntactic structures of a sentence and its translation pair in another language can be numerically represented by their Euclidean distance, calculated on the basis of the degree centralities and closeness centralities of the syntactic dependency trees of the sentences. The mean distances thus calculated for a set of translation pairs of two languages can be used as a measure of the similarity/difference between these two languages. A corpus analysis using a multi-lingual parallel corpus reveals that mean Euclidean distances thus calculated seem to reflect the typological tendencies of the differences between languages within and between language families.

1 Introduction

Sentence similarity measuring has recently attracted the attention of many researchers because it is required for various natural language applications, such as those related to question answering (De Boni and Manandhar, 2003), plagiarism detection (Alzahrani et al., 2012), and semantic searching (Farouk et al., 2018). Sentence similarity measuring has been conducted through a variety of techniques, yet the majority of them emphasize lexico-semantic similarity between sentences. Syntactic similarity measures in this context are typically employed only to augment the accuracy of semantic similarity measurement (e.g., Batanovic and Bojic, 2015; Lee et al., 2014; Ma and Suel, 2016), possibly based on the observation that the same meaning can be expressed by a variety of sentences with different syntactic structures and, inversely, that sentences with the same syntactic structure but different words can have completely different meanings. This one-to-many correspondence between sentential meaning and syntactic structures situates lexical or semantic similarity between sentences as equivalent to sentence similarity and syntactic similarity as playing a secondary, supplemental role in sentence similarity.

In spite of the success of sentence similarity measuring, its trend with an emphasis on lexico-semantic similarity should not distract us from investigating the purely syntactic similarities or differences between sentences, which remains worthy of extensive linguistics research with high quality data. The most appropriate data for this purpose is multilingual parallel corpora that consist of a large number of sentences and their translations in several languages. Since these translation pairs are semantically equivalent and the lexico-semantic differences between them are somewhat controlled, we can focus on their syntactic similarities/differences.

In analyzing syntactic similarities/differences between translation pairs in a multilingual parallel corpus, it is important to take a quantitative approach for the following reasons. First, quantifying the syntactic similarity of a given translation pair of two languages included in a multilingual parallel corpus allows it to represent the pure syntactic similarity of the translation pair, and thus to be applied to the types of natural language processing applications mentioned above as an auxiliary measure for sentence similarity. Second, and more importantly, many existing similarity analyses were subjectively conducted by individual researchers, such as those focused on

the differences and similarities in the syntactic structures of sentences in languages of different branches or families. In this context, quantitative approaches to syntactic structure can bring greater objectivity to linguistic analyses (Oya, 2014).

Addressing the need for more quantitative work in this area of linguistics, this study proposes (1) that the syntactic-structural property of a sentence can be numerically represented as the graph centralities of the dependency tree of the sentence, (2) that the difference in the syntactic structures between a sentence and its translation can be numerically represented by their Euclidean distance, and (3) that the average of the distances between a set of translation pairs can be used as a measure of similarity or difference between the two languages. To this end, this study makes a number of assumptions. First, we assume that the structural setting of the syntactic dependency structure for a sentence can be represented by two unique centrality values of the dependency relationships among the words in the sentence: degree centrality and closeness centrality. A dependency tree, which is a formalism of syntactic structure, has one unique degree centrality and one unique closeness centrality. Therefore, these two unique values can be used as unique coordinates, based on which it is possible to calculate the Euclidean distance between them. Second, we assume that the syntactic dependency structures of translation pairs are identical when the Euclidean distance between them is zero, and that their similarity is inversely proportional to the Euclidean distance between them (i.e., the more distant they are from each other, the less similar they are to each other). Since translation-pair sentences have the same meaning, the semantic difference between them is controlled as a minimum; thus, we can presume that the Euclidean distance between these syntactic structures represents the purely syntactic difference between these two sentences of a translation pair. Third, we assume that the mean Euclidean distance thus calculated between translation pairs from two languages of the same language branch (or family) is shorter than that between two languages of different branches. This assumption is based on the insight that the translation pairs of two languages of the same language branch may share similar structural settings because they are semantic equivalents. To verify this assumption, the study (1) used sentences from a multi-lingual parallel corpus that includes translation pairs of sentences from Indo-European languages such as Germanic, Romance, and Slavic as well as from non-Indo-European languages such as Chinese, Japanese, and Finnish; (2) calculated the degree centralities and closeness centralities of these sentences; (3) calculated the Euclidean distances between the translation pairs of these languages; and (4) compared these distances across languages of the same linguistic branch and of different linguistic branches.

2 Dependency grammar and typed-dependency trees

A recent trend in syntactic analysis is the emergence of numerous dependency-based frameworks, such as Word Grammar (Hudson, 2010), the Extensible Dependency Grammar (Debusmann and Kuhlmann, 2007), the Stanford Dependency (De Marneffe and Manning, 2012), and Universal Dependency (McDonald et al., 2013; Nivre et al., 2016; Tsarfaty, 2013; De Marneffe et al., 2014; Zeman, 2015). These modern developments in dependency grammar frameworks originate from Tesnière (1959). Tesnière's notion of dependency grammar can be summarized as follows (Oya, 2020): (1) each word in a sentence is dependent on another word, (2) no word in a sentence is independent, and (3) the dependency relationship between words is directed from a head to a tail. For example, Figure 1 outlines the dependency tree for the sentence "David has written this article."

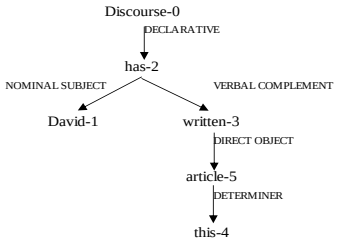

Figure 1. The dependency tree for "David has written this article."

The formalism and dependency types chosen in this study are based on Universal Dependency (UD), which makes explicit the connections among the words in a sentence.

The characteristics of such a network can be quantified in several ways based on methods from

the field of graph theory (Oya, 2014). In other words, the structural properties of networks of words in sentences can be made explicit in dependency grammar and then quantified using graph theory.

In particular, the representation of dependency relationships among words can be interpreted as a directed acyclic graph (DAG) in the UD framework. Therefore, the dependency tree for a sentence is a DAG and represents the formal syntactic property of the sentence.

3 Graph centralities and the Euclidean distances of syntactic dependency trees

The characteristics of a given graph are defined by various measures in graph theory (Freeman, 1979; Wasserman and Faust 1994). One such measure is centrality. The centrality of a node in a graph represents its relative importance within the graph. Two types of graph centrality are employed in this study: degree centrality and closeness centrality.

3.1 Degree centrality

Degree centrality is defined by the degree of a given node, that is, the number of edges a given node has (Freeman, 1979; Wasserman and Faust, 1994). The degree centrality of a node in a graph is formally represented as follows:

$$C'_D(n_i) = \frac{d(n_i)}{g-1} \quad (1)$$

where $C'_D(n_i)$ is the degree centrality of the graph, $d(n_i)$ is the degree of the ith node of the graph n, and g is the number of the nodes in the graph (Wasserman and Faust, 1994).

The degree centrality of the whole graph C_D is the sum of the maximum degree in the graph minus the degree of each of all the other nodes, divided by the largest possible sum of the maximum degree of the graph minus the degree of all the other nodes (Wasserman and Faust, 1994). The degree centrality ranges from 0 to 1:

$$C_D = \frac{\sum_{i=1}^{g}[C_D(n^*)-C_D(n_i)]}{max\sum_{i=1}^{g}[C_D(n^*)-C_D(n_i)]} \quad (2)$$

The degree centrality of a given typed-dependency tree of a sentence indicates the extent to which the words in the sentence are dependent on one particular word (Oya 2014). Thus, the degree centrality of a syntactic dependency tree can be interpreted as its flatness. Degree centrality equals 1 when a node is adjacent (connected by one edge) to all the other nodes in a graph. In terms of the dependency among words in a sentence, the degree centrality of a syntactic dependency tree is 1 when a particular word is the dependency head of all the other words in a sentence.

Degree centrality decreases as the edges connecting nodes in a graph become less concentrated on one particular node; in terms of dependency among words in a sentence, the degree centrality of a syntactic dependency tree decreases as the dependencies among words become less concentrated on one particular word in a sentence.

3.2 Closeness centrality

The distance from one node to another is represented by the number of edges between them. Freeman (1979) and Wasserman and Faust (1994) define closeness centrality as the reciprocal of the sum of the length of a path from one node to another in a graph. The closeness centrality of a graph is calculated as follows (Sabidussi, 1966; Wasserman and Faust, 1994; Beauchamp, 1965):

$$C_c(n_i)=\frac{g-1}{\sum_{j=1}^{g} d(n_i,n_j)} \quad (3)$$

where g means the number of nodes and $d(ni,nj)$ is the shortest path (geodesic distance) between the nodes n_i and n_j. Closeness centrality thus calculated can be viewed as the inverse average distance between node i and all the other nodes in the graph, ranging from 0 to 1 (Wasserman and Faust, 1994).

As in the case of degree centrality, closeness centrality equals 1 when a node is adjacent to all the other nodes in a graph. In terms of the dependency among words in a sentence, the closeness centrality of a syntactic dependency tree is 1 when one particular word is the dependency head of all the other words in a sentence.

Closeness centrality decreases as the nodes are aligned further away from each other in a graph; meanwhile, in terms of dependency among words in a sentence, closeness centrality of a dependency tree represents the extent to which its words are close to each other along its dependencies, and thus numerically indicates the embeddedness of the words in the tree; greater closeness centralities mean less embedded dependency trees (Oya, 2014).

3.3 The Euclidean distance between syntactic dependency structures

The degree centrality and closeness centrality of one syntactic dependency tree can be used as unique coordinates to indicate the structural setting of the tree, because one syntactic dependency tree contains unique degree and closeness centralities. On the basis of these coordinates of the dependency trees of translation-pair sentences, it is possible to calculate the Euclidean distance, that is, the structural similarity, between them.

The Euclidean distance is calculated as follows. Let there be two points in the Cartesian coordinates, represented as $p = (p_1, p_2,..., p_n)$ and $q = (q_1, q_2,..., q_n)$, respectively; then, the Euclidean distance from p to q (or from q to p) is calculated by the following formula:

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_1 - p_1)^2} \quad (4)$$

The Euclidean distance between a sentence and its translation pair can be calculated with their degree centralities on the x-axis and their closeness centralities on the y-axis on the assumption that degree centralities and closeness centralities are orthogonal coordinates. For example, the Euclidean distance between a sentence in one language s_1 and its translation pair in another language s_2 is calculated as follows. Suppose sentence s_1 has a degree centrality 0.22 and a closeness centrality 0.33 and the sentence s_2 has a degree centrality 0.14 and a closeness centrality 0.41. These sentences emerge as two vectors, the first of which is s_1 (0.22, 0.33) and the second is s_2 (0.14, 0.41). The Euclidean distance between them is calculated as follows:

$$d(s_1, s_2) = \sqrt{(0.22 - 0.14)^2 + (0.33 - 0.41)^2} \approx 0.113 \quad (5)$$

The Euclidean distance between a sentence and its translation pair in one language allows us to calculate the distance between the sentence and its translation pairs from other languages. For example, the Euclidean distance between sentences s_1 and s_3, which is its translation pair from a third language, can be calculated by the same procedure described above. Suppose that sentence s_1 has a degree centrality 0.22 and a closeness centrality 0.33, and sentence s_3 has a degree centrality 0.20 and a closeness centrality 0.35. Then, these sentences emerge as two vectors, the first of which is s_1 (0.22, 0.33) and the second is s_3 (0.2, 0.35). The Euclidean distance between them is calculated by the following formula below:

$$d(s_1, s_3) = \sqrt{(0.22 - 0.2)^2 + (0.33 - 0.35)^2} \approx 0.028 \quad (6)$$

Thus, the Euclidean distance between s_1 and s_2 (approximately 0.113) is greater than that between s_1 and s_3 (approximately 0.028). In other words, s_1 is closer to s_3 than to s_2 with respect to their structural settings, which are hierarchically represented by dependency among the words, and numerically by their degree centralities and closeness centralities. This means that the syntactic structure of s_1 is more similar to that of s_3 than that of s_2.

3.4 Comparisons of the Euclidean distances calculated from a parallel-corpus

Notice that the idea of Euclidean distance between a sentence from Language A and a sentence from Language B described in the previous section can further be applied collectively by contrasting the Euclidean distances between sentences from Language A and sentences from Language B and further contrasting the Euclidean distances between sentences from Language A and sentences from Language C. If it is found that the frequencies of shorter distances between Languages A and B are significantly larger than those between Language A and Language C, then it can be concluded that Language A is structurally closer to Language B than to Language C.

The overall Euclidean difference between Language A and Language B can be represented as the distribution of the frequencies of the Euclidean distances between the translation pairs of these two languages along with a certain interval. If Language A and Language B are structurally similar, then the frequencies of shorter Euclidean distances between them are expected to be higher than those of longer ones; hence, their distribution will be skewed to the left. On the other hand, if Language C is less similar to Language A than Language B, then the frequencies of shorter Euclidean distances are expected to be smaller than those between Language A and B; here, the peak of the distribution goes to the right. If the difference between these two distributions (i.e., that for Languages A and B and that for Languages

A and C) is statistically significantly large, then the structural difference between Language A and Language B is statistically significantly larger than that between Language A and Language C. (cf. Section 5.2).

The Euclidean distances between translation-pair sentences in a multilingual parallel corpus can indicate the dependency-structure settings of these languages, especially the similarity/difference between their syntactic structures. Since these translation pair sentences have the same meaning, their semantic similarity or difference is appropriately controlled for purely syntactic-structural comparisons or contrasts between these languages.

4 Corpus analysis: the Euclidean distance between dependency trees in terms of centrality measures

4.1 Purpose

As discussed in the Introduction, the purpose of this analysis is to examine the assumption that the distance between syntactic structures, or syntactic dependency trees, is expected to represent a purely syntactic similarity or difference between two languages from either the same language family or from different ones.

4.2 Data: The Parallel Universal Dependency corpus

This study used Parallel Universal Dependency (PUD) treebanks as the data for calculating the Euclidean distance between translation pairs based on the degree centrality and closeness centrality of their syntactic dependency trees.

PUD treebanks were created for the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to UDs. PUD treebanks contain 19,000 sentences from 19 different languages (Arabic, Chinese, Czech, English, Finnish, French, German, Hindi, Indonesian, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Thai, and Turkish). The subcorpus of PUD treebanks for each language contains 1,000 sentences, in a fixed order across languages. The sentences are aligned one-to-one, although some sentences are translated into two sentences. Of the 1,000 sentences, 750 are translated from English texts, and the remaining 250 sentences are translated from German, French, Italian, or Spanish, which were translated into English and then translated into other languages (their ID numbers indicate the original language). The translation was done by professional translators and annotated with morphological and syntactic tags by Google. They were then converted by UD community members to match UD Version 2 guidelines. For further details on PUD treebanks, refer to the UD webpage (https://universaldependencies.org/).

In summary, PUD treebanks is a set of subcorpora such that (1) each subcorpus of a language contains 1,000 sentences of that language, (2) these 1,000 sentences have their translation counterparts in 18 other languages (1,000 multiplied by 18 equals 18,000 translation pairs for one language), and (3) PUD treebanks ultimately contain 342,000 translation pairs (18,000 multiplied by 19).

4.3 Method

The word count, degree centrality, and closeness centrality of each sentence in PUD treebanks was calculated by an original Ruby script. Another Ruby script was created to calculate the Euclidean distance of the syntactic dependency trees of all the translation pairs in PUD treebanks.

Then, a spreadsheet application was used to calculate the frequencies of Euclidean distances (cf. Section 3.4) between the translation pairs in all the unique combinations of 19 languages (19 languages multiplied by 18 other languages minus non-unique combinations equal 171 combinations).

Translation pairs from different languages show different distributions of Euclidean distances. For example, 1,000 translation pairs exist between English and Japanese in PUD treebanks, and the 1,000 Euclidean distances between them are distributed from 0 to 0.85, with the most frequent one 0.1 with an interval of 0.01 (65 pairs). There are also 1,000 translation pairs between English and German, and their 1,000 Euclidean distances are distributed from 0 to 0.7, with the most frequent 0.04 (90 pairs).

The distributions of these frequencies in each of these unique combinations were compared by a Wilcoxon's signed-rank test, in order to check whether the difference between these distributions

was wider than that caused by chance. This test was chosen because the frequencies of the Euclidean distance between translation pairs did not seem to be normally distributed.

For reasons of space, we focus here on our comparisons of the Euclidean distance between English and Japanese, and the Euclidean distance between English and other languages (18 comparisons overall). The English and Japanese languages were chosen as the focus of comparison (Language A and B in Section 3.4) because they do not belong to the same language family; English belongs to the Indo-European family, while Japanese does not; thus, comparing them provides a starting point for comparisons of other language pairs, which will be conducted in future studies.

5 Results

5.1 The Euclidean distances between different languages in PUD treebanks

The first row in Table 1 summarizes the descriptive statistics *of* the Euclidean distances among the syntactic dependency trees of all the translation pairs in the PUD treebanks in terms of degree centralities and closeness centralities. All the other rows show the Euclidean distances between the syntactic dependency trees of each language and those of all the other 18 languages in the PUD treebanks in terms of their degree centralities and closeness centralities.

Finnish (a non-Indo-European language) has the longest mean Euclidean distance from all the other languages (0.139) while Portuguese (a Romance language, Indo-European family) has the shortest (0.100). The mean Euclidean distances of 9 out of 11 Indo-European languages in PUD treebanks are less than 0.11, while those of 6 out of 8 non-Indo-European languages in PUD treebanks are more than 0.11.

Finnish also has the largest SD (0.139) while German has the smallest (0.075). The SDs of 9 out of 11 Indo-European languages in the PUD treebanks are less than 0.8, and those of 7 out of 8 non-Indo-European languages are more than 0.8.

	Euclidian distance						
	Mean	Median	Mode	Max.	Min.	SD	N
All	0.112	0.095	0	1.005	0	0.083	342,000
Arabic	0.123	0.109	0	1.002	0	0.082	18,000
Chinese	0.118	0.103	0	1.002	0	0.082	18,000
Czech	0.110	0.093	0	1.002	0	0.083	18,000
English	0.102	0.086	0	0.853	0	0.077	18,000
Finnish	0.139	0.115	0	0.869	0	0.105	18,000
French	0.104	0.088	0	0.835	0	0.079	18,000
German	0.104	0.089	0	0.869	0	0.075	18,000
Hindi	0.115	0.101	0	1.005	0	0.079	18,000
Indonesian	0.103	0.088	0	0.793	0	0.078	18,000
Italian	0.102	0.087	0	0.835	0	0.078	18,000
Japanese	0.121	0.106	0	0.856	0	0.081	18,000
Korean	0.137	0.121	0	1.002	0	0.091	18,000
Polish	0.107	0.092	0	1.002	0	0.079	18,000
Portuguese	0.100	0.084	0	0.856	0	0.078	18,000
Russian	0.102	0.087	0	0.865	0	0.077	18,000
Spanish	0.101	0.085	0	0.775	0	0.078	18,000
Swedish	0.105	0.087	0	1.002	0	0.082	18,000
Thai	0.110	0.093	0	1.005	0	0.086	18,000
Turkish	0.125	0.109	0	1.002	0	0.087	18,000

Table 1. The mean Euclidean distances of the syntactic dependency trees of each language to those of all the other languages.

5.2 Comparing the Euclidean distances between different language pairs

This section reports the comparisons of the frequencies of the Euclidean distances between English and Japanese, and those between English and other languages, all in PUD treebanks. These comparisons are intended to show the structural differences between languages in terms of the different distributions of their Euclidean distances based on their degree centralities and closeness centralities.

Of all the 18 comparisons, only one pair (English-Japanese and English-Swedish; Swedish is the Language C in Section 3.4) showed a significant difference between the distributions of the frequencies of the distances, and three others show slightly significant differences (those between English–Japanese and English–German, English–Polish, and English–Spanish).

Figure 2 describes one of the instances in which the two distributions show a significant difference (English-Japanese and English-Swedish).

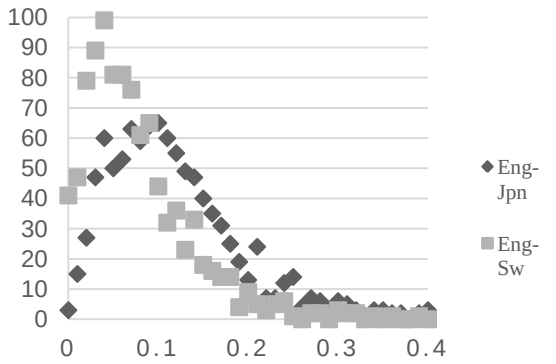

Figure 2. The frequencies of the Euclidean distances between the English sentences and their Swedish translations (Eng-Sw; N = 1,000), and those between the English sentences and their Japanese translations (Eng-Jpn; N = 1,000); x axis: Euclidean distances (interval: 0.01; max: 0.4); y axis: frequencies

A Wilcoxon Signed-Ranks Test indicated that the number of short Euclidean distances between English and Swedish translation pairs was statistically significantly larger than the number of short Euclidean distances between English and Japanese translation pairs ($Z = 2.01$, $p = 0.044$). This means that a significantly larger number of translation pairs of English and Swedish in the PUD treebanks are structurally closer to each other than the English and Japanese pairs.

Figure 3 describes one of the instances in which the two distributions do not show any significant difference; one distribution includes the frequencies of the Euclidean distances between the English sentences and their Japanese translations, and the other those between the English sentences and their Chinese translations.

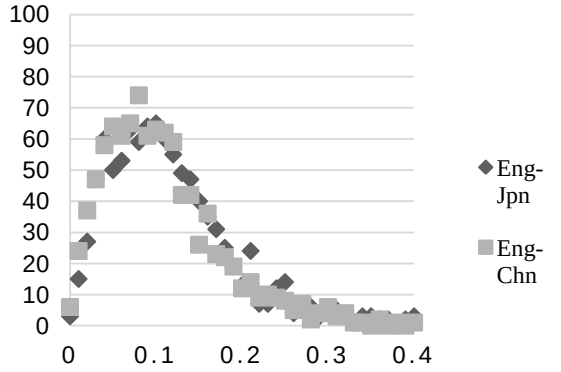

Figure 3. The frequencies of the Euclidean distances between the English sentences and their Chinese translations (Eng-Chn; N = 1,000), and those between the English sentences and their Japanese translations (Eng-Jpn; N = 1,000); x axis: Euclidean distances (interval: 0.01; max: 0.4); y axis: frequencies

A Wilcoxon Signed-Ranks Test indicated that the frequency of short Euclidean distances between English and Japanese translation pairs was not larger than the frequency of short Euclidean distances between English and Chinese translation pairs ($Z = 0.5$, $p = 0.617$). This means that there is no statistically significant difference between the distribution of frequencies of English-Japanese Euclidean distances and those of English-Chinese Euclidean distances.

6 Discussion

The results of the comparisons of the Euclidean distances between the two pairs of languages does not seem to contradict the assumption that the distance between the syntactic dependency trees calculated by the two graph centrality measures represents a purely syntactic difference between two languages. It was found that the mean Euclidean distances of syntactic dependency trees of each language to all the others are divided approximately into two groups. The first group has shorter Euclidean distances and includes Indo-European languages, and the second group has longer Euclidean distances and includes non-Indo-European languages. It was also found that the distribution of frequencies of the Euclidean distances between English (a Germanic language, Indo-European family) and Japanese (a non-Indo-European language) shows no significant difference with that between English and Chinese (a non-Indo-European language), a slightly significant difference with that between English and Spanish (a Romance language, Indo-European family), and a significant difference with that between English and Swedish (a Germanic language, Indo-European family). These results might lead us to assume that the Euclidean distances among syntactic dependency trees calculated above seem to represent the structural similarity of the languages of the same language family/branch and structural difference/diversity of languages of different language families/branches. To verify this assumption, further comparisons must be made between more language pairs in the PUD treebanks or other parallel corpus data with a

larger variety of languages, which is our research goal for future studies.

7 Conclusion

This study has proposed the idea that the difference of syntactic structures of a sentence and its translation pair in another language can be numerically represented by their Euclidean distance, calculated on the basis of the degree centralities and closeness centralities of the syntactic dependency trees of sentences, and that the mean distances thus calculated for a set of translation pairs of two languages can be used as a measure to show similarity/difference between these two languages. The corpus analysis using a multi-lingual parallel corpus revealed that along with some interesting properties of the graph centrality measures of syntactic dependency trees, mean Euclidean distances between the syntactic dependency trees of translation-pair sentences of a variety of languages seem to reveal their typological tendencies. Further comparisons are needed between more language pairs in PUD treebanks or other multi-lingual parallel corpus data.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 20K00583.

References

Salha M. Alzahrani, Naomie Salim, and Ajith Abraham. 2012. Understanding plagiarism linguistic patterns, textual features, and detection methods. IEEE Transactions on Systems, Man, and Cybernetics, 42(2):133-149. https://doi.org/10.1109/TSMCC.2011.2134847.

Vuk Batanovic and Dragan Bojic. 2015. Using part-of-speech tags as deep-syntax indicators in determining short-text semantic similarity. Computer Science and Information Systems, 12(1):1-31. https://doi.org/10.2298/CSIS131127082B.

Murray A. Beauchamp. 1965. An improved index of centrality. Behavioral Science, 10:161-163.

Marco De Boni and Suresh Manandhar. 2003. The use of sentence similarity as a semantic relevance metric for question answering. Proceedings of the AAAI Symposium on New Directions in Question Answering, SS-03-07.

Marie-Catherine de Marneffe and Christopher D. Manning. 2012. Stanford Typed Dependency Manual Revised for the Stanford Parser v.2.0.4. Retrieved May 30, 2020 from http://nlp.stanford.edu/software/dependencies_manual.pdf.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. Proceedings of LREC14.

Ralph Debusmann and Marco Kuhlmann. 2007. Dependency Grammar: Classification and exploration. Project report (CHORUS, SFB 378). Retrieved May 30, 2020 from http://www.ps.uni-saarland.de/~rade/papers/sfb.pdf

Mamdouh Farouk, Mitsuru Ishizuka, and Danushka Bollegala. 2018. Graph matching based semantic search engine. Proceedings of 12th International Conference on Metadata and Semantics Research, Cyprus. https://doi.org/10.1007/978-3-030-14401-2_8.

Linton C. Freeman. 1979. Centrality in social networks. Social Networks, 1:215-239.

Richard Hudson. 2010. An Introduction to Word Grammar. Cambridge University Press, Cambridge.

Ming Che Lee, Jia Wei Chang, Tung Cheng Hsieh. 2014. A grammar-based semantic similarity algorithm for natural language sentences. The Scientific World Journal. https://doi.org/10.1155/2014/437162.

Weicheng Ma and Torsten Suel. 2016. Structural sentence similarity estimation for short texts. 29th International Florida Artificial Intelligence Research Society Conference. Key Largo, United States. 232-237.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 92-97.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. Proceedings of the Tenth International Conference on Language Resources and Evaluation, 1659-1666.

Masanori Oya. 2014. A Study of Syntactic Typed-Dependency Trees for English and Japanese and Graph-centrality Measures [Doctoral dissertation] Waseda University, Tokyo, Japan.

Masanori Oya. 2020. Structural divergence between root elements in English-Japanese translation pairs. Journal of Global Japanese Studies, 12:107-126.

Gert Sabidussi. 1966. The centrality index of a graph. Psychometrika, 31: 581−603.

Lucien Tesnière. 1959. Éléments de syntaxe structural. Klincksieck, Paris.

Reut Tsarfaty. 2013. A unified morpho-syntactic scheme of Stanford dependencies. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 578-584.

Stanley Wasserman and Katherine A. Faust. 1994. *Social network analysis.* Cambridge University Press.

Daniel Zeman. 2015. Slavic Languages in Universal Dependencies. In Slovko 2015: Natural Language Processing, Corpus Linguistics, E-learning. Bratislava, Slovakia.