Evaluating WordNet's Efficacy in Representing Human Conception and Categorization of Semantics

Mengqing Deng (m.deng@mail.utoronto.ca)
University of Toronto

Jack (Zhaoxue) Li (zhaoxue.li@mail.utoronto.ca)
University of Toronto

Abstract

WordNet, originally conceived as a model of human semantic organization, has evolved into a crucial resource for natural language processing (NLP) research and applications. Despite its widespread usage, it remains unclear whether WordNet continues to accurately reflect human cognitive processes. In this study, we investigate the relationship between human performance on concept recognition tasks involving word pairs from the Semantic Priming Project (SPP) and various similarity metrics derived from WordNet for the same pairs. We hypothesize that higher WordNet similarity ratings correspond to improved human concept recognition in terms of speed and accuracy. To test this hypothesis, we conduct significance tests on reaction times and accuracy scores for word pairs exhibiting different similarity ratings. Our results demonstrate that word pairs with higher similarity ratings exhibit a significant advantage in concept recognition tasks compared to those with lower ratings. In other words, higher similarity or relatedness ratings are associated with enhanced human performance in recognition tasks. These findings indicate that WordNet continues to capture certain aspects of human semantic organization, particularly in the context of categorization tasks, and reveal potential connections between NLP methodologies and cognitive functions related to concept recognition.

Keywords: WordNet; SSP; Categorization; Concept recognition; Semantic relatedness; Reaction Time

Introduction and Related Works

WordNet

WordNet, initially developed in 1985 at the Cognitive Science Laboratory of Princeton University under the guidance of psychologist George A. Miller and later directed by linguist Christiane Fellbaum, was originally conceived as a model to represent human semantic organization (Miller, 1995). The linguistically-oriented database offers a robust structure, rich semantic relations, and multiple classification systems, which facilitates natural language processing (NLP) tasks such as word sense disambiguation, information retrieval, and semantic similarity measurement. These attributes make it an ideal tool for understanding and measuring lexical meaning as well as semantic relatedness in humans' natural language. Over time, this well-established large-scale lexical database has evolved into an invaluable resource for researchers and practitioners in linguistics and NLP, and its original objective has been overshadowed (Fellbaum et al, 2010). Despite its widespread adoption and utility in the NLP domain, it remains an open question whether WordNet continues to accurately reflect human cognitive processes as it was initially intended. Despite this shift in focus, we hypothesize that WordNet can still reveal some insights into how humans categorize concepts.

At the core of WordNet's structure are synsets, which are groups of semantically equivalent words or phrases, also known as *synonyms*. Synsets represent a single distinct concept or meaning, and they are interconnected through various types of semantic relations, the most fundamental of which are *hypernyms* (broader or more general concepts) and *hyponyms* (narrower or more specific concepts). Other relations include *meronyms* (parts), *holonyms* (wholes), *antonyms* (opposites), and *troponyms* (manner-based sub-events). These semantic relationships between synsets allow for the creation of hierarchical network structures with respect to various taxonomies, which resembles a node graph or a tree, that can be navigated with computing tools (Fellbaum, 2005).

WordNet comprises several distinct classification systems, namely the noun, verb, adjective, and adverb hierarchies, each organized independently but interconnected through the semantic relations mentioned earlier. The noun taxonomy is the most extensive, containing top-level concepts such as the *entity*, *physical object*, and *event*. It branches out into more specific categories and subcategories as it moves down the hierarchy. The verb taxonomy revolves around actions and processes, while adjectives and adverbs pertain to properties and modifiers, respectively. These classification systems enable researchers to better understand and measure lexical meaning and semantic relatedness in the English language.

In this study, we aim to investigate the relationship between WordNet and human conception of semantics. Various approaches have been developed for measuring semantic similarity between words using WordNet (Agirre et al., 2009; Budanitsky & Hirst, 2006). In the methodologies section, we will delve into the details of these approaches as we process the data and derive different similarity metrics from WordNet.

Priming Effect

We posit that semantic similarity is crucial in determining the way lexicons are represented in human cognition. Numerous studies in the fields of psychology and linguistics have indicated that the priming effect substantially influences human perception and recognition of pairs of lexical or semantic information within a brief time frame (Meyer & Schvaneveldt, 1971; Seidenberg et al., 1984). When a participant is exposed to a *target* word preceded by a *prime* word, the greater

the relatedness between the two words, the more accessible the information embedded in the target word becomes for the participant. Nevertheless, the organization and similarity of concepts within WordNet may not correspond to human categorization processes as suggested by Exemplar and Prototype theories (Murphy, 2016).

In this study, we propose that presenting humans with pairs of concepts that have a higher similarity rating derived from WordNet, will result in faster and/or more accurate recognition of the concepts. To study human performance in recognition tasks of pairs of words, we utilize the data collected in the following project.

The Semantic Priming Project (SPP)

In the study conducted by Hutchison et al. (2013) on the Semantic Priming Project (SPP), the authors designed experiments that involved two word recognition tasks, namely the Speeded Naming Task (NT) and Lexical Decision Tasks (LDT), to investigate the semantic priming effect in humans. The study comprises a dataset of 1,661 target words, and results were derived from both tasks. A total of 768 participants were recruited from four major universities in the United States, and the dataset includes demographic information and descriptive characteristics for each participant. As part of the research, participants completed sections of the Woodcock-Johnson reading battery (Woodcock et al., 2001), three attentional control tasks, and a circadian rhythm measure.

Our primary focus is on the NT datasets, as its experimental design aligns more closely with our research objectives. Within the NT trials, reaction time and accuracy records were obtained by initially presenting the participants with either a related or unrelated priming word, followed by an inter-stimulus interval (ISI) (also referred to as stimulus onset asynchrony, SOA) period, and subsequently displaying the target word. Participants were then instructed to enunciate the target word as rapidly and accurately as possible.

Materials and Methodologies

To enhance our data analysis and extract pertinent information, we employed various data processing techniques on the data sourced from SPP and WordNet. The code and processed datasets used in our study are accessible via the following repository on GitHub: `https://GitHub.com/jack-li-2000/COG403_final`.

In this project, we utilized the Natural Language Toolkit (NLTK) library (Bird et al., 2009b) in Python to access WordNet. Our work was conducted with the following versions of the resources: NLTK 3.8.1, WordNet 3.0, and Python 3.10.10.

Data Retrieval and Pre-Processing

First and foremost, it is essential to note that the interactive database, initially provided by the SPP, is no longer maintained or accessible. As a result, we rely on the static data available at `https://www.montana.edu/attmemlab/spp.html` (Hutchison, 2010). A significant portion of the dashboard features discussed in the paper (Hutchison et al., 2013) are now unavailable, including the relatedness scores between words, as well as the specific algorithms used to quantify them in the original study. Consequently, we cannot reference or compare these with our calculated ratings.

Moreover, the determined sense and part-of-speech (POS) for each word of choice in their study, initially obtainable through the interactive database, are no longer accessible as well. This has further complicated our research and increased its difficulty. We were compelled to employ several prediction and disambiguation algorithms to determine the characteristic attributes of the words in the experiment. However, these algorithms lack 100% accuracy and cannot compare to the manually controlled standards and ground truth labels in the original study, which are no longer available.

In the SPP, the relationships between the prime and target word pairs presented to participants fell into several categories (Hutchison et al., 2013). First, there were word pairs with the same POS, such as synonyms (e.g., *frigid–cold*), antonyms (e.g., *hot–cold*), category coordinates (e.g., *table–chair*), and category superordinate relations (e.g., *dog–animal*). These pairs are well-suited for investigating their relatedness through the semantic network structure of WordNet. The other 8 categories encompass forward/backward phrasal associates (FPA/BPA, e.g., *help–wanted/wanted–help)*, perceptual properties (e.g., *canary–yellow*), functional properties (e.g., *broom-sweep*), script relations (e.g., *restaurant–wine*), instrument relations (e.g., *broom–floor*), actions (e.g., *scrub-dishes*), associated properties (e.g., *deep–dark*), and unclassified relations (e.g., *mouse–cheese*). Most of these word pairs commonly co-occur in corpora, forming phrases or collocations or possessing functional or causal relationships, making it challenging to directly determine their relatedness using WordNet. Since the category information is no longer accessible from the SPP, we will introduce an alternative algorithm to partition the data into two parts for WordNet and collocation analysis, generate relatedness scores that optimally match their appropriate strength, and ensure they are as meaningful as possible.

The pertinent fields from the NT data for our investigation include prime and target words, naming accuracy, reaction time, and ISI, involving 413564 rows of trial data. The comprehensive dataset has been cleaned and isolated for use in statistical testing and analysis, with the specific process to be discussed in greater detail in the later sections.

Despite employing imperfect algorithms that introduced a certain degree of noise into our data for analysis, the significance of our findings remains undiminished. These methods managed to attain satisfying functionality, and although the introduced noise was not negligible, it did not substantially impact the overall conclusions of our research and still led to noteworthy discoveries.

Processing in WordNet

Word Sense Disambiguation (WSD) It stands to reason that each lexical entry in natural language can have multiple possible meanings. In WordNet, the majority of words are attached with more than one synset - some common words could even have up to dozens of them. Thus, to measure the relatedness between two words, it is essential to disambiguate their meanings and determine the specific synsets before applying the formulae of similarity ratings.

Upon consolidating the data, we found that there were a total of 257 participants in the NT experiment, with each participant completing approximately 1,600 trials. There were 6,911 unique prime-target pairs, with each pair being tested around 60±30 times. We decided to determine the synset for each unique pair and store the results in a hash table for future reference, thereby simplifying and avoiding redundant calculations. Since the POS of the words in the data is unknown, it is not easy to lemmatize them. Consequently, we only stripped the leading and trailing white space and converted each word to lowercase before obtaining all synsets for the word using the NLTK. We then iterated through the two sets of synsets for each word pair to find the two synsets that have the shortest connecting path in the Wordnet network structure, and fixed them for use in the subsequent similarity calculation formula.

In our WSD task, due to the absence of context information where the words to be disambiguated only appear in pairs, it is impractical to apply conventional methods such as Lesk or BERT. Our proposed algorithm demonstrates superior performance compared to the widely used Most Frequent Sense (MFS) approach in this scenario.

Similarity Ratings from WordNet To utilize WordNet for computing similarity, we examined popular and well-regarded general similarity algorithms from the past several decades (Agirre et al., 2009; Budanitsky & Hirst, 2006), including Path Similarity, Wu-Palmer Similarity (Wu & Palmer, 1994), Leacock-Chodorow Similarity (Leacock & Chodorow, 1998), Resnik Similarity (Resnik, 1995), Jiang-Conrath Similarity (Jiang & Conrath, 1997), and Lin Similarity (Lin, 1997). These algorithms are embedded within the NLTK library, and their specific formulas can be found in the aforementioned original papers or in WordNet documentation (Goodman, 2020). The first three are taxonomy-based metrics, which primarily depend on the structure of the taxonomy network and the number of edges in some specific paths between synset nodes. The latter three introduce the concept of Information Content (IC) and leverage the weight of words in a corpus to assist in computing similarity.

Processing in Corpora

Collocation Relatedness Ratings by N-Gram We utilized four popular corpora provided by NLTK (Bird et al., 2009a) to train an n-gram model (Brown et al., 1992) - in our case, specifically a bigram model - for predicting the probability of a target word occurring after a given prime word. These corpora include the Reuters Corpus (Lewis, 1997), the Brown Corpus (Francis & Kucera, 1979), the Project Gutenberg Corpus (Hart, 1971), and the Webtext Corpus (Liu & Curran, 2006). By providing a prime and a target word, we can obtain a probability score ranging from 0 to 1, which serves as a measure of the words' relatedness within the given corpus. We then further compute a weighted average based on the word frequency weights of the different corpora.

Upon obtaining these collocation ratings, we can use them to partition the dataset into two segments. If a collocation rating is non-zero, it implies that the word pair has appeared as consecutive words in the corpus, making it less likely that they belong to the same POS categories. Consequently, their WordNet similarity may not be very meaningful during data analysis. Therefore, these word pairs are separated into a collocation dataset, distinct from the same POS categories word pairs suitable for WordNet analysis.

Visualization of Ratings Distribution

The visual representation of the distribution for each rating, calculated across the entire NT dataset, can be observed in Figure 11 located in the appendix.

Further Collation and Split up for Analysis

ISI In order for WordNet and SPP data to be suitable for statistical testing and plotting, some further steps need to be completed. Using the Pandas library from python, perform an inner merge on the two datasets and group the data into isi = 50ms and isi = 1050ms. The data is separated based on isi since we believe the amount of delay between the priming word and target word for the word tasks from SPP will affect the results of the significance test.

Threshold For comparisons with Wu-Palmer similarity, choose a threshold to separate the isi50 and isi1050 into two additional sets; for this analysis, we choose the threshold to be 0.50. This means there will be four total datasets: isi_50_0, isi_50_1, isi_1050_0, isi_1050_1. The 0 at the end represents less than 0.50 similarity rating for the word pairs and the 1 at the end representing the opposite. A higher similarity score indicates a more similar relationship between the meanings of the words. The Wu-Palmer Similarity rating is a linear metric that ranges from 0 to 1; knowing this, we choose the threshold - point of separate for the data sets - to be 0.50 arbitrarily.

Isolation For the collocation data, the process is slightly different. Since we may find that higher similarity rated word pairs have a significant effect on human reaction time and accuracy, it is considered an unwanted effect and we isolate for only the low similarity rating word pairs. All other steps are the same with the exception that we use the entirety of the collocation data as one distribution and use the low Wu-Palmer similarity rated word pairs as the other distribution for the t-test.

Removing Duplicate Word Pairs

In the SPP and WordNet data, there are many duplicate word pairs which is a problem for plotting. Duplicate word pairs can result in issues with plotting, specifically duplicate similarity ratings resulting in multiple datapoints for the same x value. To prevent this, take the average of each word pair rather than having a single datapoint for each.

Statistic Tests Applied

The rational for using the t-test as our method of computing significance is because we are separating the combined dataset into two datasets based on a filter, not randomly. By filtering the data, the distributions of the resultant datasets are likely altered to where their means and standard deviations are different from before. Using the t-test, we can compare the descriptive statistics of the two sets and test if they are significantly different. From the results, we determine if the difference between higher similarity rating and lower similarity rating data sets is due to random chance or if it is unlikely to be a coincidence. Furthermore, the p-value from the t-test is determined separate for isi50 and isi1050.

Results

	Reaction Time	Accuracy
isi_50ms	1.884367e-13	1.228786e-09
isi_1050ms	1.478007e-11	1.098564e-13

Figure 1: Reported p-values for reaction time and accuracy for both isi from t-test.

	RT_mean	RT_std	acc_mean	acc_std
less related	555.848528	190.212895	0.984758	0.122514
more related	549.453492	186.403051	0.988017	0.108811

Figure 2: Description of distribution of 50 isi data separated by more similar vs less similar.

	RT_mean	RT_std	acc_mean	acc_std
less related	545.557447	204.618815	0.984980	0.121632
more related	539.241982	199.779348	0.988894	0.104798

Figure 3: Description of distribution of 1050 isi data separated by more similar vs less similar.

Reaction Time With the WordNet metric of Wu-Palmer similarity with a score higher than 0.50 inclusive for synonyms and lower than 0.50 for unrelated words, we find that there is a significant difference in mean values in both the 1050 ms data and the 50 ms data sets (figures 2,3). From the 50 ms data set, the reported p-value is 1.884367e-13 (figure 1) which is statistically significant with very little chance for the difference in mean to be random. For the 1050 ms data set, the conclusion is the same with a p-value of 1.478007e-11. In the following plots, we see that the reaction time vs similarity rating plot both have a downward trend with the 50 ms plot having a more drastic decrease (figures 7, 9).

Accuracy Accuracy between Wu-Palmer score of below 0.50 and above 0.50 were also both significant with a p-value much lower than 0.05 with the results being 1.228786e-09 for 50 ms and 1.098564e-13 for 1050 ms (figure 1). Descriptively, both the accuracy mean for Wu-Palmer similarity of below and above 0.50 are very close to 1 with similar variances. The mean differences for accuracy are not large, but undeniably significant (figures 2,3,8,10).

	Reaction Time	Accuracy
isi_50ms	5.622441e-10	4.551689e-04
isi_1050ms	8.900303e-10	6.810777e-07

Figure 4: Reported p-values for reaction time and accuracy for both isi from t-test.

	RT_mean	RT_std	acc_mean	acc_std
less related	555.848528	190.212895	0.984758	0.122514
more related	541.131921	185.500618	0.990066	0.099182

Figure 5: Description of distribution of 50 isi data separated by collocation vs not collocation.

	RT_mean	RT_std	acc_mean	acc_std
less related	545.557447	204.618815	0.984980	0.121632
more related	529.655571	190.444584	0.992552	0.085986

Figure 6: Description of distribution of 1050 isi data separated by collocation vs not collocation.

Collocation Analysis

With collocation, we see very similar results as Wu-Palmer similarity. Reaction time and accuracy means are significantly different between collocated word pairs for both 50ms isi and 1050ms isi (figure 4 gives the actual p-values). Similar to Wu-Palmer similarity, the means for reaction time and ac-

curacy match our hypothesis; no collocation (labeled less related) have a higher mean reaction time and lower mean percentage accuracy when compared with collocation (labeled more related) for both isi of 50ms and 1050ms (figure 5,6).

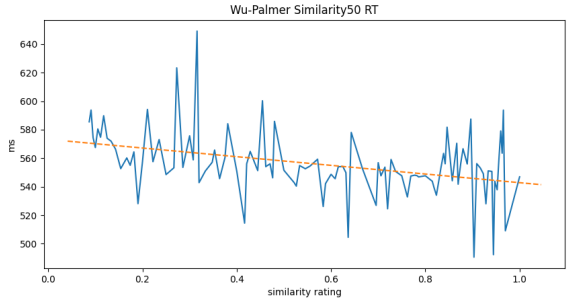

Figure 7: Reaction time vs Wu-Palmer similarity rating for 50 ms.

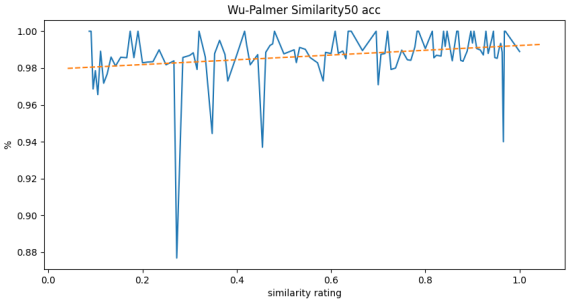

Figure 8: Accuracy vs Wu-Palmer similarity rating for 50 ms.

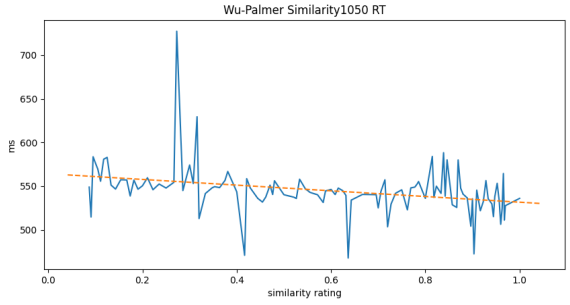

Figure 9: Reaction time vs Wu-Palmer similarity rating for 1050 ms.

Figure 10: Accuracy vs Wu-Palmer similarity rating for 50 ms.

Discussion

Reaction time and accuracy in relation to Wu-Palmer similarity score performed as expected for both 50 ms and 1050 ms groups.

Reaction time and accuracy in relation to collocation performed as expected for both 50ms and 1050ms groups.

Wu-Palmer Similarity metric

We chose to use the Wu-Palmer similarity rating because the distribution of the ratings is very linear. Other ratings we have seen before, like Path similarity, are non-linear and heavily left skewed. Since they all strive to measure similarity in WordNet, we use the Wu-Palmer metric for its ease of use and compatibility with plotting.

Dataset Separation Threshold

For the choice of Wu-Palmer similarity score threshold, we chose 0.50 as an arbitrary choice. It is equally viable to have the upper section be 0.8 and above and lower section be 0.2 and below; in fact, selecting 0.8 and 0.2 would likely increase the chance of rejecting the null. However, our results are significant with the choice of separating by 0.50 similarity rating and there is no reason to go beyond this choice.

Collocation Separation

For collocation, we decided to choose between complete lack of collocation between the word pairs and collocated word pairs. This decision is because even low ratings of collocation between the pair gives the target word some familiarity. An example of a collocated pair with a low rating is ‘last chance’, while this pair may not occur together as frequently allowing it a high rating, this pair is still very notable.

Future Steps

In the current study, we perform analysis on two major concepts relating words; synonyms and collocation. We can extend our analysis further with other Synsets from WordNet and perform feature analysis to determine the weighting of each feature on human performance of word tasks. For example, it is conceivable that antonyms, hyponyms, and more are compared to test which concept matters more in improving human performance in the SPP word tasks.

Conclusion

In the conclusion, we reject the null hypothesis. We successfully determined that Wu-Palmer similarity rating in the WordNet data structure significantly reflects the reaction time and accuracy performance of subjects at both the 1050 ms and 50 ms SOA groups. While the difference between the sets are not large, we may certain look toward this result for insight and draw strong conclusions. Since the Wu-Palmer similarity rating reflects our depiction of concept similarity in WordNet, we imagine that more similar concepts in WordNet is correlated with better performance in concept recognition in people. Furthermore, we may suggest that collocated word pairs from WordNet also indicated better performance of reaction time and accuracy for word tasks in participants.

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009, June). A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 19–27). Boulder, Colorado: Association for Computational Linguistics.

Bird, S., Klein, E., & Loper, E. (2009a). Accessing Text Corpora and lexical Resources. In *Natural language processing with python – Analyzing Text with the Natural Language Toolkit.* O'Reilly Media Inc.

Bird, S., Klein, E., & Loper, E. (2009b). *Natural language processing with Python: Analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Brown, P. F., Della Pietra, V. J., deSouza, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-Based *n*-gram Models of Natural Language. *Computational Linguistics*, *18*(4), 467–480.

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, *32*(1), 13–47. doi: 10.1162/coli.2006.32.1.13

Fellbaum, C. (2005). WordNet and wordnets. In *Encyclopedia of Language and Linguistics* (pp. 2–665).

Fellbaum, C., Poli, R., Healy, M., & Kameas, A. (2010). WordNet. *Theory and Applications of Ontology: Computer Applications*.

Francis, W. N., & Kucera, H. (1979). *Brown Corpus Manual* (Tech. Rep.). Department of Linguistics, Brown University, Providence, Rhode Island, US.

Goodman, M. W. (2020). *Wn 0.9.3 documentation.* https://wn.readthedocs.io/en/latest/index.html.

Hart, M. S. (1971). *Project Gutenberg.*

Hutchison, K., Balota, D., Neely, J., Cortese, M., & et al Cohen-shikora, E. (2013). The Semantic Priming Project. *Behavior Research Methods*, *45*, 1099–1114.

Hutchison, K. A. (2010). *Semantic Priming Project - Attention & Memory Lab — Montana State University.* https://www.montana.edu/attmemlab/spp.html.

Jiang, J. J., & Conrath, D. W. (1997, August). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the 10th Research on Computational Linguistics International Conference* (pp. 19–33). Taipei, Taiwan: The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Leacock, C., & Chodorow, M. (1998, January). Combining Local Context and WordNet Similarity for Word Sense Identification. In *WordNet: An Electronic Lexical Database* (Vol. 49, p. 265).

Lewis, D. D. (1997). *Reuters-21578 Text Categorization Test Collection, Distribution 1.0.*

Lin, D. (1997, July). Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 64–71). Madrid, Spain: Association for Computational Linguistics. doi: 10.3115/976909.979626

Liu, V., & Curran, J. R. (2006). Web Text Corpus for Natural Language Processing. In D. McCarthy & S. Wintner (Eds.), *EACL.* The Association for Computer Linguistics.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*, 227–234. doi: 10.1037/h0031564

Miller, G. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*, 39–41.

Murphy, G. (2016). Is there an exemplar theory of concepts? *Psychonomic Bulletin & Review*, *23*, 1035–1042.

Resnik, P. (1995, November). *Using Information Content to Evaluate Semantic Similarity in a Taxonomy* (No. arXiv:cmp-lg/9511007). arXiv.

Seidenberg, M. S., Waters, G. S., Sanders, M., & Langer, P. (1984, July). Pre- and postlexical loci of contextual effects on word recognition. *Memory & Cognition*, *12*(4), 315–328. doi: 10.3758/BF03198291

Woodcock, R. W., Mather, N., McGrew, K. S., & Wendling, B. J. (2001). Woodcock-Johnson III tests of cognitive abilities.

Wu, Z., & Palmer, M. (1994, June). Verb Semantics and Lexical Selection. In *32nd Annual Meeting of the Association for Computational Linguistics* (pp. 133–138). Las Cruces, New Mexico, USA: Association for Computational Linguistics. doi: 10.3115/981732.981751

Appendix

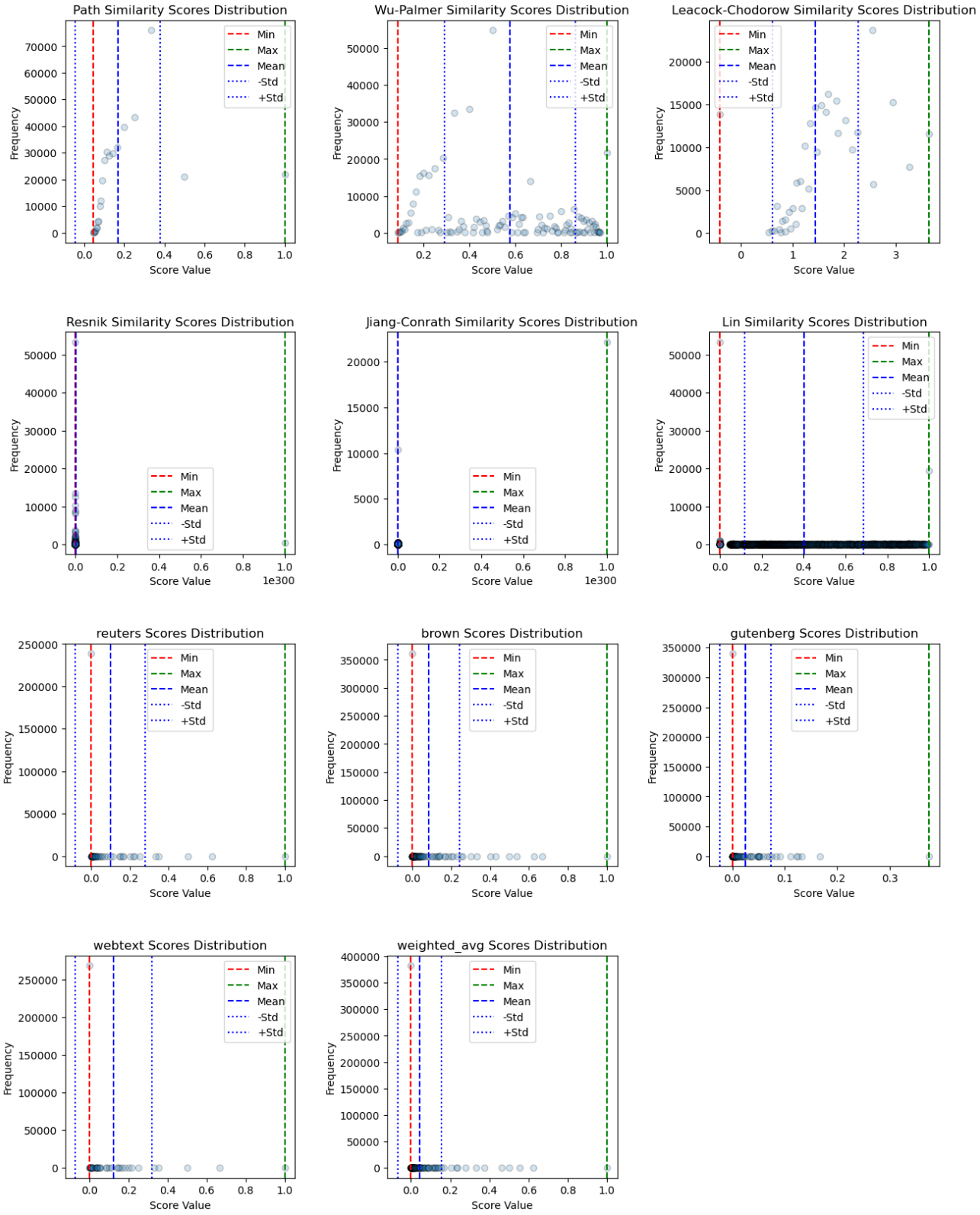

Figure 11: Distribution of all Ratings in NT Dataset