International Journal of Data Science and Analysis
2022; 8(1): 18-22
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20220801.13
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Using Human Intelligence to Test the Impact of Popular Preprocessing Steps and Feature Extraction in the Analysis of Human Language

W. Randolph Ford[1], Ingrid G. Farreras[2, *]

[1]Department of Data Science, Harrisburg University of Science and Technology, Harrisburg, United States of America

[2]Department of Psychology, Hood College, Frederick, United States of America

Email address:

rford@harrisburgu.edu (W. R. Ford), farreras@hood.edu (I. G. Farreras)

[*]Corresponding author

To cite this article:

W. Randolph Ford, Ingrid G. Farreras. Using Human Intelligence to Test the Impact of Popular Preprocessing Steps and Feature Extraction in the Analysis of Human Language. *International Journal of Data Science and Analysis.* Vol. 8, No. 1, 2022, pp. 18-22.
doi: 10.11648/j.ijdsa.20220801.13

Received: January 10, 2022; **Accepted**: February 5, 2022; **Published**: February 16, 2022

Abstract: More than half a century has passed since Chomsky's theory of language acquisition, Green and colleagues' first natural language processor Baseball, and the Brown Corpus creation. Throughout the early decades, many believed that once computers became powerful enough, the development of A.I. systems that could understand and interact with humans using our natural languages would quickly follow. Since then, Moore's Law has basically held; computer storage and performance has kept pace with our imaginations. And yet, 60 years later, even with these dramatic advances in computer technology, we still face major challenges in using computers to understand human language. The authors suggest that these same exponential increases in computational power have led current efforts to rely too much on techniques designed to exploit raw computational power and, in so doing, efforts have been diverted from advancing and applying the theoretical study of language to the task. In support of this view, the authors provide empirical evidence exposing the limitations of techniques – such as n-gram extraction – used to pre-process language. In addition, the authors conducted an analysis comparing three leading natural-language processing question-answering systems to human performance, and found that human subjects far outperformed all question answering-systems tested. The authors conclude by advocating for efforts to discover new approaches that use computational power to support linguistic and cognitive approaches to natural language understanding, as opposed to current techniques founded on patterns of word frequency.

Keywords: Natural Language Processing, NLP, N-gram, Phrase-tructure Parsing, AllenNLP, DeepPavlov.ai, BERT

1. Introduction

The decade between 1957 and 1967 was a remarkable time in the development of our understanding of human language. Noam Chomsky [1] presented his theory of language acquisition, Bert Green and colleagues [2] wrote Baseball, the first natural language processor, and Kučera and Francis [3], in creating the Brown Corpus, began to describe the nature of word frequency in human language. As a result, in the late 1960s, most believed that by interweaving these three distinct approaches to the understanding of language – the theoretical, the simulation, and the computational approaches –, that computers would soon not only understand human language, but that we would also be interacting with computers via human conversation. This, however, is not what has transpired over the last half century.

The current state of technology has yet to meet these expectations. Moreover, and specifically in the last 10 years, new approaches to language analysis seem to have left behind the ever-growing academic understanding of theoretical linguistics. Perhaps as a result of the impact of Moore's Law [4], researchers have turned instead to novel approaches and techniques that rely heavily on raw computational power and applied mathematical methods in pursuit of processing human language by computer. When considering the accelerated development of technology over

the last half century [5], this is not surprising, for nothing in academic study has come close to keeping abreast with the advance of technology. And, in the realm of this technology, these approaches were not just the path of least resistance, but not taking advantage of developing resources would have been foolish. Panesar [6] accurately and thoroughly traces these paths that have led us to our current state of natural language understanding.

Nevertheless, by favoring methods that rely so much on computational power, perhaps some of the most critical components of language understanding are now being systematically overlooked. In fact, and perhaps as a result, several leaders in the field of natural language understanding have begun to note the incomplete nature of our current approaches. Manning [7] described that language is not about what words mean, but about what people mean, and added that although computers can now recognize words with astounding accuracy, they still cannot understand what people mean. Not long after, Mikolov [8] concurred, pointing out how researchers still do not know how to model the long-term memory all humans rely on to make sense of language. Unlike computers, humans use existing memory to sort, relate, reduce, and even reject new information, with a result that can be far from the original input and not always dependent on a perceptual stream. The human mind can, and frequently does, create new content based on cognitive processes and emotions that are independent of external stimuli. And most recently, Dunietz, Burnham, Bharadwaj, Rambow, Chu-Carroll, and Ferrucci [9] demonstrated how existing approaches fall short of real comprehension, of even being able to define what a valid measure of comprehension should entail.

The beginning of this current trend of focusing on computational power to unravel human language can arguably be traced all the way back to the Brown Corpus. This was one of the very first efforts where computers were used to process human language from the perspective of word frequency of use. One of the most interesting revelations from the Brown Corpus was proof that a very small subset of words accounted for the vast majority of our communication. While this concept was first uncovered manually in the 1930s by Dolch [10], the Brown Corpus provided large-scale computational proof of it 30 years later [3]. And today, in the Corpus of Contemporary American English (COCA) [11], the most frequently used 100 words account for about 42% of the one billion words appearing in the corpus.

It is not just that a small percentage of words account for the vast majority of those used in human communication, but that the profile of parts of speech of those frequently-used words is strikingly different from the profile of parts of speech of communication itself [12]. In analyzing the COCA from a parts-of-speech perspective, only five of those 100 most frequently used words – "people", "back", "way", "time", and "years" – are classified as common nouns and account for less than 2% of the total frequency of those 100 most frequently-used words. Prepositions, on the other hand, account for 23%, while determiners account for almost 25% of the frequency of the 100 most frequently-used words.

The key issue here is that unique nouns occurring more than 12 times in the COCA are coming from a pool of 40,000 unique words, adjectives from a pool of 30,000 unique words, and verbs from a pool of 20,000 unique words. Prepositions, on the other hand, come from fewer than 150 unique words, determiners come from 11 unique words, and pronouns come from 10 unique words. Hudson [13] reported that 37% of words used in English communication are nouns. The underlying SVO structure of the English language predicts the frequent occurrence of nouns as subject and object, both direct and indirect. Additionally, the occurrence of a preposition should, in most cases, indicate the presence of a noun.

Does this then mean that techniques relying on word frequency in preprocessing steps to computational analyses, such as n-grams, may be transforming human communication away from the probability of understanding it? There is no arguing against the value of n-gram models in providing a wonderful process for predicting the next word in communication for help with text messaging and other such applications. But does this functionality ensure that statistical approaches to natural language processing, such as n-grams, can also be used for understanding natural language? Dunietz, Burnham, Bharadwaj, Rambow, Chu-Carroll, and Ferrucci [9] demonstrated how these techniques are significantly lacking when compared to human performance on comprehension tasks. The underlying assumption in using these techniques is that, even though they are transforming human language away from its naturally occurring form, at some point – with increasing computational power – we will still reach understanding. Ford and Berkeley III [14], however, in an analysis of the most frequently-used n-grams starting with the word "influence" appearing in the COCA, noted that about one third of these n-gram phrases ended with a preposition, another third ended with either a possessive pronoun or an article, and in most n-grams, it was impossible to determine if "influence" was being used as a noun or a verb.

In fact, the analyses that led to two recent patents, Ford and Berkeley III [14], and Ford, Berkeley III, and Newman [19] uncovered that the general trend in frequency of word use decreases as one traces this dependent variable from the beginning to the end of a linguistic phrase. For example, for noun phrases, words used as articles and possessive pronouns have a higher frequency of use than words used as adjectives, which have a higher frequency of use than words used as nouns. These two patents successfully demonstrate a process for rapidly breaking sentences into linguistic phrases by identifying patterns of words based on decreasing patterns of word frequency.

Given the above findings, there is little doubt that many of the pre-processing steps used in modern approaches to deep learning, such as n-gram extraction, are transforming natural language in the process. It seems that separating

language based on collections of words based on the frequency of those words occurring together, results in pulling language patterns apart in a way that obscures meaning instead of exposing it. But how can this be proven? This study attempts to answer this question by reintroducing the human ability, acknowledged by Manning [7] and Mikolov [8] to be missing, back into the process. Specifically, can humans still understand language once preprocessing transformations used in modern language analysis are applied?

The authors conducted this experiment to determine empirically how much and what kind of information needs to be present in a reading passage for people to be able to make sense of it. The authors compared the reading comprehension that results from information provided by n-gram analyses vs. information provided by phrase-structure parsing phases, to determine which of the two approaches most closely approximates the degree of reading comprehension obtained when individuals have access to a full text. The prediction was that phrase-structure parsing phrases would result in greater reading comprehension than n-gram analyses would, despite n-gram analyses being one of the most popular tools used today to analyze language.

Furthermore, the authors also processed the reading-comprehension material, in their original state, though three popular A.I. question-answering systems – AllenNLP [15], BERT [16], and DeepPavlov.ai [17] – and compared the results to the human performance (predicted to be superior).

2. Method

Participants consisted of 95 traditional, undergraduate students enrolled in multiple undergraduate psychology courses offered at a private, Mid-Atlantic liberal arts college. Some faculty members from those courses offered the students extra credit for participating in the experiment. Subsequent to obtaining approval from the college's Institutional Review Board, and signed informed consent and publication forms, participants were randomly assigned to one of three between-subjects conditions. All participants received four reading passages either in intact form (Full Text condition), in phrase-structure parsing phrase form (Phrase-Structure Parsing Phrase condition), or in n-gram form (N-gram condition).

To address some of the dataset concerns outlined by Dunietz, Burnham, Bharadwaj, Rambow, Chu-Carroll, and Ferrucci [9], the reading passages and questions all came from psychometrically sound, publicly available (https://www.crackssat.com/isee/reading) Upper Level Reading Comprehension tests related to history, science, literature, and contemporary life that form part of the Independent School Entrance Examination (ISEE), administered by the Educational Records Bureau to students applying for grades 9-12 admission. For the N-gram condition, the authors ran an n-gram analysis on the 20-passage corpus (approximately 10,000 words in length) that created 5,080 unique n-grams with a total frequency of 17,130 for those n-grams that occurred more than once. The n-grams created ranged from five words down to single words in length, and those were the ones used for the n-gram condition (beginning with the largest n-grams).

For the Phrase-Structure Parsing Phrase condition, the authors used a computer-based phrase-structure parser that was part of a legacy NLP system [18], and extracted the highest frequency n-grams and phrase-structure parsing phrases for each for the two experimental conditions, and randomly selected four reading passages among all of the ones available (number of words' range for each passage: 265-340), together with the four to six four-answer multiple choice questions that accompanied each passage. Participants were allowed all the time they needed to complete reading all of the material in their condition and answering the reading comprehension questions (a total of 20).

The correction of the three question-answering systems' responses was determined by two judges who were instructed to make their determinations using the most liberal criteria possible. The judges independently agreed on solution correctness in all cases.

3. Results

3.1. First Analysis

For the first analysis, the authors coded the total number of correct answers out of the 20 reading comprehension questions for each participant. A one-way between-subjects ANOVA compared the average correct answers for each condition, and found a statistically significant difference across all reading conditions: $F(2,92)=35.68$, $p=0.001$, $\eta_p^2=0.437$, power=1.00. A planned post-hoc LSD was used to determine the nature of the differences among the conditions, and found that participants in the Phrase-Structure Parsing Phrase condition (range: 10-18; $SD=2.03$) understood equally well as participants in the Full Text condition (range: 11-19; $SD=2.09$) ($p=0.065$), but participants in the N-gram condition (range: 6-16; $SD=2.37$) understood significantly less than participants in both the Full Text and the Phrase-Structure Parsing Phrase conditions ($p=0.001$ and 0.001, respectively, with a Bonferroni-corrected alpha level of 0.016 (i.e., 0.05/3)).

3.2. Second Analysis

For the second analysis, the authors compared the human participants' performance to the three popular NLP question-answering systems. BERT answered 45% of the questions (9 of 20) correctly, AllenNLP system answered 40% of the questions (8 of 20) correctly, and DeepPavlov.ai answered 25% of the questions (5 of 20) correctly. At 78% accuracy, human participants far outperformed all three question-answering systems (see Figure, with 95% confidence intervals), confirming that statistical approaches that employ word-frequency as an underlying technique are inefficient at capturing meaning.

3.3. Figure

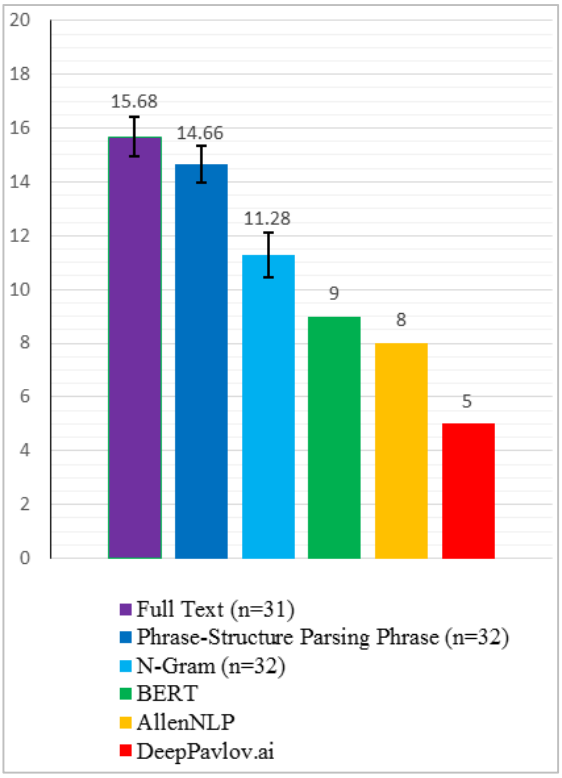

Figure 1. *Mean Number of Correct Reading Comprehension Answers (out of 20) by Condition.*

4. Discussion

This study was one demonstration that current, mathematical techniques – specifically n-grams – employed to analyze language are altering language to such a degree that it cannot be understood by people. The authors compared the comprehension of 95 adults reading full texts vs. phrases extracted via phrase-structure parsing from these texts vs. via high-frequency n-gram analyses. The results demonstrated that those participants relying on phrases extracted through n-gram analyses were the least likely to understand the material, and that three leading NLP question-answering systems "understood" the material even less.

Why did they all perform worse than the Phrase-Structure Parsing Phrase condition? The authors suggest it is because n-grams change communication. Word frequency influences the process of n-gram creation, so word types – such as nouns, verbs, and adjectives – occur at a lower frequency in n-grams. When comparing the unique words among the three conditions, the authors discovered that there were 327 word types that occurred less frequently, or not at all, in the n-gram phrase condition when compared with the other two conditions, and all but five of those word types were either nouns, verbs, or adjectives.

N-gram frequency analyses have tried to define the syntactic structure of human language understanding by expanding the analysis of word frequency to frequently-occurring phrases in the identification of noun and verb groupings. It makes intuitive sense that tracking frequently-used phrases would yield meaningful results, so what could be simpler than running an n-gram analysis on any corpus of text to get a list of important noun phrases? However, when one needs to employ extensive post-processing to derive value from a technique – a common course of action in such analyses – the approach becomes suspect, and transforms language in a manner that further obstructs understanding. This does not apply only to n-gram analyses, but to other statistically-derived approaches to understanding language such as bag of words, one hot encoding, sentiment analyses, and others that change the parts-of-speech profile of language to the point where understanding is jeopardized.

The authors believe that current deep learning techniques in natural language processing are limited because they are based on methods where, in the first step, elements are extracted from language to be used in processing. If we accept that human language has evolved under the counteracting forces of our need to efficiently express communication while employing unpredictability and novelty to hold the attention of others, then *any* technique based on element extraction, whether words or n-grams, will have its limits. Attempting to extract the meaning of language from words as opposed to phrases that are combined to form sentences will yield limited results, given that meaning is contained at a higher hierarchical level than words.

5. Conclusion

If we accept that the question-answering systems tested in this analysis are representative of the state-of-the-art of current computational systems, then we are looking at the highest results possible coming out of these deep learning techniques. It may well be that some future modification to these techniques will ultimately overcome the difficulties that arise from trying to understand language at a word level. But, as things now stand, they perform below the level of human participants. Perhaps this is an indication that it is time to reconsider our current mathematical approaches to feature-extraction natural language processing, and instead use computational power to support linguistic and cognitive approaches that better explain how the brain functions.

A starting point would be to analyze the correct and incorrect responses that come out of these systems to determine where they align and where they do not align with human reasoning and memory. Standardized reading comprehension passages and multiple choice questions serve as an excellent foundation for this task in that their psychometric properties inform us of how they align with human performance. The authors hope that this approach will lead to more fruitful and promising possibilities for simulating how humans process and understand language.

Acknowledgements

We thank Dr. George Dimitoglou (Professor and Chair of the Computer Science Department of Hood College) and Dr.

Kevin Purcell (Associate Professor of Data Science and Director of the Masters in Analytics Program at Harrisburg University), for their suggestions on a draft of this manuscript.

References

[1] Chomsky, N. (1957). Syntactic Structures. Boston, MA: Mouton de Gruyter.

[2] Green, B., Wolf, A., Chomsky, C., & Laughery, K. (1961). Baseball: An automatic question answerer. Western joint IRE-AIEE-ACM computer conference, 19, 219-224.

[3] Kučera, H., & Francis, W. N. (1967). Computational analysis of present-day American English. Providence, RI: Brown University Press.

[4] Moore, G. E. (1965, April 19). Cramming more components onto integrated circuits. Electronics, 38 (8), 114 ff.

[5] Roser, M., & Ritchie, H. (2017). Technological progress. https://ourworldindata.org/technological-progress.

[6] Panesar, K. (2020). Conversational artificial intelligence: Demystifying statistical vs linguistic NLP solutions. Journal of Computer-Assisted Linguistic Research, 4, 47-79. http://hdl.handle.net/10454/18121.

[7] Manning, C. D. (2016, June 23). Language is communication; Texts are knowledge. The Future of Artificial Intelligence Conference, Stanford University, Stanford, CA. https://www.vimeo.com/173057086.

[8] Mikolov, T. (2018, August 25). When shall we achieve human-level AI? Human-Level AI Conference, Prague, Czech Republic. https://www.slideslive.com/38910040/when-shall-we-achieve-humanlevel-ai.

[9] Dunietz, J., Burnham, G., Bharadwaj, A., Rambow, O., Chu-Carroll, J., & Ferrucci, D. (2020). To test machine comprehension, start by defining comprehension. Proceedings of the 58th annual meeting of the Association for Computational Linguistics, 7839-7859. https://www.aclweb.org/anthology/2020.acl-main.pdf.

[10] Dolch, E. W. (1936). A basic sight vocabulary. The Elementary School Journal, 36 (6), 456-460.

[11] Davies, M. (2020-). The Corpus of Contemporary American English (COCA): 1 billion words, 1990-present. https://www.english-corpora.org/coca.

[12] Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41 (4), 977-990.

[13] Hudson, R. (1994). About 37% of word tokens are nouns. Language, 70 (2), 331-339. https://doi.org/10.2307/415831.

[14] Ford, W. R., & Berkeley III, A. R. Understanding natural language using tumbling-frequency phrase chain parsing. U.S. Patent No. 10,783,330, September 22, 2020. http://www.freepatentsonline.com/y2020/0125641.html.

[15] Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. H. S., Peters, M. E., Schmitz, M., & Zettlemoyer, L. (2018). AllenNLP: A deep semantic natural language processing platform. Proceedings of the workshop for NLP Open Source Software (NLP-OSS), 1-6. https://www.aclweb.org/anthology/W18-2501.pdf.

[16] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 4171-4186. https://doi.org/10.18653/v1/N19-1423.

[17] Burtsev, M., Seliverstov, A., Airapetyan, R., Arkhipov, M., Baymurzina, D., Bushkov, N., Gureenkova, O., Khakhulin, T., Kuratov, Y., Kuznetsov, D., Litinsky, A., Logacheva, V., Lymar, A., Malykh, V., Petrov, M., Polulyakh, V., Pugachev, L., Sorokin, A., Vikhreva, M., & Zaynutdinov, M. (2018). DeepPavlov: Open-source library for dialogue systems. Proceedings of the 56th annual meeting of the ACL: System Demonstrations, 122-127. https://doi.org/10.18653/v1/P18-4021.

[18] Ford, W. R. Multi-stage pattern reduction for natural language. U.S. Patent No. 7, 599, 831, October 6, 2009. http://www.freepatentsonline.com/7599831.html.

[19] Ford, W. R. & Berkeley, A. R., Newman, M. A. Understanding natural language using split-phrase tumbling frequency phrase-chain parsing. U.S. Patent No. 11,055,487, July 6, 2021. http://www.freepatentsonline.com/11055487.html.