The need for an increased understanding of natural language processing is universal in scope and potential utility. Businesses examine linguistic data to gain insights in meeting the needs of customers. Models of language classification contribute to the design of assistive technologies and teaching methods for individuals that may experience difficulties in reading, writing, visual or auditory comprehension, or in speech production. Evidence suggests that human communication shares characteristics in terms of neural processing regardless of language. Grammatical and semantic properties of words are an essential aspect of all linguistic models. However, many traditional approaches result in statistical models that are overfitted or too highly specified, resulting in a lack of applicability across language or domain. On the other hand, linguistic models aiming for generalization are often overly broad, also resulting low practical functionality (Nivre, 2015). Recent increases in data accessibility, computational power, as well as innovative strategies in textual data manipulation has led to exponential growth and advancement in this topic area. This has contributed to increased demand for methodological development to improve upon existing theory and approaches.

The purpose of this project is to examine the complex process of word classification and discrimination using evolutionary algorithms in order to test and assess potential opportunities for optimization. As this project is exploratory, the eventual experimental design will consider and may incorporate multiple parameters associated with word classification and word discrimination. Schmidt et al. applied differential evolution in hypertuning parameters themselves, which may support the use of a multimodal design. That said, focus on fine-tuning a single procedural step may be preferable, allowing for more trials and greater detail. While libraries and open-source tools have made experimentation more accessible for researchers and firms, many are not sufficient in providing real-time analysis or supporting use in distributed systems (Landset, Khoshgoftaar, Richter, & Hasanin, 2015). One potential measure to assess the quality of lemmatization is applying a genetic approach to optimize speed and memory used in the operations, a considerable limitation in the analysis of large corpora. Adjusting decision trees by applying genetic algorithms using a random forest model may increase performance or contribute to improvements in word classification, producing stop-words, spell-checking, parts of speech, and semantic meaning most closely approximating natural language processing. Alternatively, Lee, Lim, & Ahn suggest that graph-directed models may improve on tree-based models. Gleim et al. (2019) assessed lemmatization comparing fine-grained parts of speech as opposed to coarser groups, a parameter with potential for optimization using genetic algorithms.

Baayen, Chuang, Shafaei-Bajestan, & Blevins suggests word composition and semantic meaning is secondary to linear discriminative learning, emphasizing phonemes, particularly trigrams, over morpheme based theories (2019). Baayen et al. model is based on matrices composed of simple linear correlational mapping between words with impressive results, greatly improving on traditional predictive capabilities. The matrices applied in Baayen et al.’s model may provide an opportunity to reduce dimensionality and result in a more minimal representation (2019). The researchers address that improving on this method with neural network comparison is promising, and aligns with connectivist linguistic theoretical approaches.

References

Baayen, R. H., Chuang, Y. Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. *Complexity*, 2019.

Buchanan, E. M., De Deyne, S., & Montefinese, M. (2019, July 24). A practical primer on processing semantic property norm data. <https://doi.org/10.31219/osf.io/qb8en>

Bungum, L., & Gamback, B. (2010, November). Evolutionary algorithms in natural language processing. In *Norwegian Artificial Intelligence Symposium, Gjøvik* (Vol. 22).

Gleim, R., Eger, S., Mehler, A., Uslu, T., Hemati, W., Lücking, A., ... & Hoenen, A. (2019). Practitioner’s view: A comparison and a survey of lemmatization and morphological tagging in German and Latin. *Journal of Language Modelling*, *7*(1), 1-52.

Landset, S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data*, *2*(1), 24.

Lee, J. H., Lim, S., & Ahn, C. W. (2019). Automotive ECU Data-Based Driver’s Propensity Learning Using Evolutionary Random Forest. *IEEE Access*, *7*, 51899-51906

Nivre, J. (2015). Towards a Universal Grammar for Natural Language Processing. *CICLing*.

Schmidt, M., Safarani, S., Gastinger, J., Jacobs, T., Nicolas, S., & Schülke, A. (2019). On the Performance of Differential Evolution for Hyperparameter Tuning. *arXiv preprint arXiv:1904.06960*.