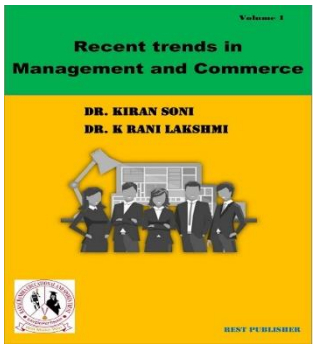
Recent trends in Management and Commerce

Vol: 1(1), 2019

REST Publisher; ISBN No: 978-81-936097-6-7

Website: http://restpublisher.com/book-series/rmc/



An Introduction Natural Language Processing

*** K Prabhakaran Anuradha**

SST College of Arts and Commerce, Maharashtra, India

*Corresponding Author Email: anuradhaprabhakaran@sstcollege.edu.in

Abstract: *The study of the way native human dialects and computer systems interact is known as natural language processing. Cognitive languages require the use of NLP, which has connections to artificial intelligence and computer engineering. In artificial intelligence, natural language processing, or NLP, includes all aspects of learning, comprehending, and creating natural human dialects. This eliminates the need for humans to use their own dialects when managing spoken and written natural speech on machines. Natural language processing, also referred to as "computational linguistics," combines semantics and grammar, to help computers comprehend human speech and writing and infer meaning from what is said. Neural networks and computer programming are combined in this area to create software that can translate languages reasonably accurately. Voice recognition is a branch of this topic that enables a computer to fully comprehend what you are saying. introduces the concept of contemporary NLP systems as well as Natural Language Processing (NLP) in general. The medical informatics generalist with little experience in NLP and/or cutting-edge technology is the target audience for this advice. We outline the typical NLP sub problems in this wide area and chart the development of NLP in light of this. We provide an overview of the key achievements in therapeutic NLP study. Following a brief introduction to common machine learning approaches for addressing various NLP sub problems, we'll discuss the architecture of current NLP frameworks and provide an overview of the Apache Foundation's Unstructured Information Management framework. Finally, we look at possible NLP directions in the future.*

Keywords: *Word2vec, Skip-gram model, natural language processing*

1. INTRODUCTION

Audit NLP, or natural language processing, is the study, comprehension, and creation of natural human communication. This eliminates the need for programs to actually use human language in order to handle spoken and recorded human language. "Natural language processing" is a field of vocabulary, technology, and machine learning which focuses on the way to program machines to handle and analyze enormous quantities of natural language input. It is an investigation of how machines and human speech interact. The goal of the "artificial intelligence" (AI) subfield of natural language processing (NLP) is to teach computers to understand written and spoken English. One of the simplest and most easy web applications of NLP is in email filters. Spam filters, which recognize specific words or phrases that indicate a communication is spam, are where it all began. NLP first appeared in the 1950s at the nexus between speech and artificial intelligence. At first, textual data retrieval (IR), which quickly indexes and finds massive amounts of text using extremely flexible statistics-based techniques, was distinguished from natural language processing (NLP): An good introduction to IR is given by Manning et al. NLP and IR have, however, started to merge somewhat over time. The mental toolkits of researchers and developers must be significantly expanded because NLP currently draws from a wide range of disciplines. The two main tasks for product review mining are the collection of product feature terms and their emotional evaluation [59]. It mainly refers to the use of instruments like machine learning for natural language processing [60], computational linguistics, fact sheets, and various other tools that rapidly recognize and extract important information from texts [61]. In essence, this refers to the extraction of high-quality data using NLP and machine learning methods from disorganized texts [31,62]. The foundation of natural language processing is figuring out how to store words in a way that computers can comprehend. One-hot representation is the most natural way to characterize word vectors, but this strong and sparse representation is extremely susceptible to the "curse of dimensionality." With this technique, the machine is unable to recall word semantic data or determine the level of relevance between word vectors. Fortunately, in order to completely address the issues raised above, To represent language, Hinton (Hinton, 1986) suggested Distributed Interpretations. One of them in their final days Word2Vec, was first developed by Tomas Mikolov et al. of Google [63]. In 2013, it was proposed to use the intermediary

portion of the model's parameters to convert the natural language lexicon from the singular form to a phrase vector model depicted by a fixed-length dense vector. The correlation among word vectors can be used to depict the semantic and logical connections between words. Essentially, this algorithm predicts words using the fundamental elements of a neural network. (Figure 1). In their educational contexts, the words present in their context as well as the target words found in the text have both been used to anticipate the words. In Word2Vec, it is straightforward to combine hundreds of dimensions to quickly create a distance matrix [63]. The distributed semantics hypothesis, which holds that words with comparable meanings commonly occur in the same context, is reflected in this semantic subspace [64]. The continuous bag-of-words (CBOW) and skip-gram methods are most commonly used when creating Word2Vec models. (Figure 2). The CBOW method uses surrounding words to predict the actual word, whereas skip-gram tries to predict phrases in a space of size c to the current word [64]. In actuality, skip-gram models frequently outperform larger data sets [64]. Therefore, a skip-gram model could be used for the investigation in this paper.

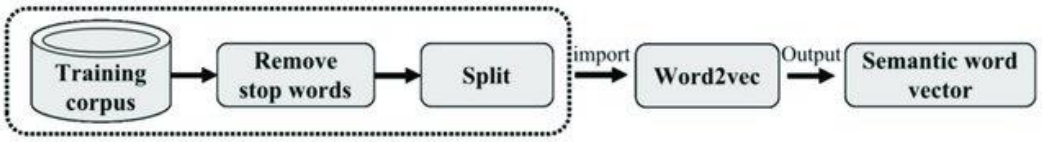


FIGURE 1. Corpus preprocessing and Word2Vec

Word2Vec: In its most basic form, a training corpus is a collection of writings used to develop algorithms for autonomous machine learning. Word2vec can be used with any writing technique. The only things that are taught are the connections between the words that are used in the text. However, not all teaching companies are created equal. The Word2vec method is employed in natural language processing to enhance text categorization. This is frequently employed in the early stages of modeling. The model is basically trained in order to comprehend the context of a word. Word2Vec is a connectionist NLP tool that converts each word's semantic meaning into numerical values that are shown as vectors with an orientation and length that show where each word is in relation to another object, such as a point, in n-dimensional space. Mikolov, Chen, Corrado, and Dean created the Word2Vec technique in 2013. It builds superior word vector models from enormous data sets using basic neural networks. The outcome is a 300-dimensional sentence vector that depicts the semantic value of the words as decided by the sentence's context. Word2Vec searches the text being studied for linked word meanings using a thin neural network that can represent the human brain. (Kiefer, 2019). The method is pragmatic and semantic because grammar and syntax are not taken into account; rather, A word's semantic definition is based on its context. In contrast to open coding, which is a summary method, embedding rules are nongenerative. They are drawn from either a response's text to determine the passage's practical meaning. A useful code must have two qualities in order to be effective: (1) it must accurately represent the content of the text it is compression algorithm, providing a high-level "summative" meaning; and (2) it must have semantic depth, encoding significant semantic "essence-capturing" that obscures the particular significance of a passage (Saldann a, 2016). The average slope between the input vectors of a term and the velocity estimates of all other terms in the phrase is a simple statistic to evaluate a term's representativeness of a sentence. The majority of other words' vectors will probably be closest to those of common, commonly used words. The word vectors with the greatest degree of isolation from other word vectors will therefore have the most precise definitions. There is a relationship between word meaning and word vector dimension. (Schakel & Wilson, 2015). A quick method to determine a word's semantic complexity is to look at its word vector magnitude. depth and accuracy (inverse average angle distance) (word vector magnitude), the two essential elements of high-quality codes, were created and utilized to create the Score= (angular vector distance)$^{-1}$ ×magnitudej; to give every word in a text a number; A large figure indicates an important code. Each application of the method requires a different tuning of the j parameter. It will be necessary to use a higher j number for a lengthier passage with ambiguous word meanings. The best codes for this corpus of interviews were generated by J = 3. We chose these specific NLP techniques for two reasons: First off, each represents a recent development in the area of natural language processing. Second, each makes use of an alternative methodology—statistical (Topic Modeling) verses neural/connectionist (Word2- Vec)—providing a wider sampling of NLP methods than if we had used one method.

2. SKIP-GRAM MODEL

In order to calculate the likelihood that X word will be the one you are looking for in that context, the skip-gram objective function adds the log probabilities of the words that are immediately on either side of the target word W (t). One method for determining the words that are most frequently linked with a phrase is the skip-gram. The context phrase for a particular target word can be predicted using skip-gram. The formula is CBOW's polar opposite. A word2vec model known as the "Skip-Gram" model was recently mentioned in a piece written for

readers who are new to NLP, but I was unable to comprehend a word of it. If someone could give a more thorough explanation of it and how it operates, that would be useful. Please, if at all feasible, share any relevant notebooks. NLP can assist patients and health consumers searching for details on a specific illness or treatment. Through machine learning to comprehend questions, which can then improve access to pertinent data, through the examination of the themes presented in an article as well as the terms used in the document, tailored to their data needs and health literacy levels. Natural language processing methods offer a way to convert unstructured text into data in a format that can be processed by computers across all of these NLP use cases. enabling software programs to process data efficiently, giving users simple access to the raw textual information, and allowing humans to communicate using recognizable natural language. The use of NLP in biology and human health is briefly explained in the part that follows.

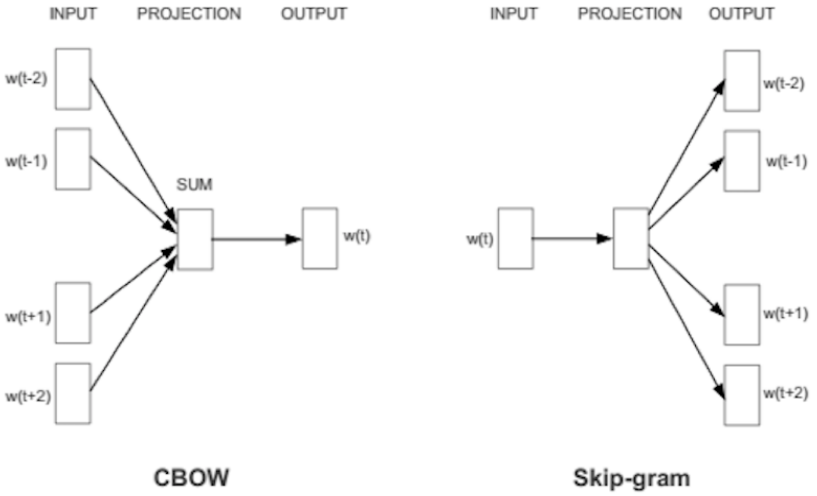


FIGURE 2. The skip-gram model

3. NATURAL LANGUAGE PROCESSING (NLP)

4.

Informatics research on natural language processing is presently very active and fascinating. The word "NLP" is frequently employed to refer to a collection of techniques for handling unstructured text, despite the techniques themselves differing greatly in how much linguistic knowledge is used. Some methods require little to no linguistic knowledge and only the presence of terms in the text. The only linguistic knowledge necessary is an understanding of the components of words, and These strategies frequently rely on key phrases or a bag-of-words approach. A engine for searching that retrieves papers based on the presence of a particular word arrangement in a collection is one example of a method that only employs words, despite the fact that the words in the documents themselves are not used in the search that are returned might not be connected to one another. Another example is a machine learning method that creates a statistical model with words as traits without taking their relationships into account. This section concentrates on additional NLP methods with a more complex understanding of language. These more complex linguistic techniques generally aim to identify the entire or to understand some of the significance of the pertinent material in texts, as well as the partial grammatical or grammatical structure of texts. Natural language is extremely large, unrestrictive, and ambiguous, which caused two issues when traditional parsing techniques that only used abstract, hand-crafted rules were used: Since formal grammars describe the connections between words and speech parts like nouns, verbs, and adjectives primarily on syntax, NLP ultimately needs to be able to extract meaning (or "seman tics") from the text. Even though this kind of writing is understandable by humans, the extremely telegraphic writing of in-hospital progress reports and "ungrammatical" conversation (in medical environments) are handled very poorly by handwritten rules. An increase in IR tools and the integration of IR technologies into relational databases can be attributed to the post-Google interest in IR. Before data mining tools, statistical algorithms also underwent commoditization. The availability of numerous instruments in a package distinguishes common software for analysis. By using a graphical metaphor, a user can frequently construct a pipeline without writing. High value in relation to price: some offerings may even be freeware. High user friendliness and ease of learning: online documentation/tutorials are highly approachable for the non-specialist, concentrating on when and how to use a specific tool rather than its underlying mathematical principles. On the other hand, NLP toolkits and UIMA continue to be targeted at experienced coders and have expensive commercial products.

4. CONCLUSION

On the other hand, NLP toolkits and UIMA continue to be targeted at experienced coders and have expensive commercial products. In the event that general-purpose NLP becomes commoditized, Best-of-breed options would have a higher success rate. The standard will likely be set once more by analytics vendors, who will carry on the efforts of biomedical informatics researchers to come up with novel solutions to the problem of processing complicated biomedical language in the various contexts in which it is used. Researchers can use NLP to judge the potential accuracy of the generated codes after using more conventional qualitative methods. The frequent reference of food in the NLP findings of this study attests to the validity of the nutrition-related codes and open coding codes. NLP is a technique for codebook verification. Researchers could use NLP to generate codes based on the findings rather than accessing the coding system first. To maximize the usefulness of NLP results, researchers might want to compare their proposed interview questions to NLP methods.

REFERENCES

[1]. Leeson, William, Adam Resnick, Daniel Alexander, and John Rovers. "Natural language processing (Nlp) in qualitative public health research: a proof of concept study." *International Journal of Qualitative Methods* 18 (2019): 1609406919887021.

[2]. Feldman, Susan. "NLP meets the Jabberwocky: Natural language processing in information retrieval." *ONLINE-WESTON THEN WILTON*- 23 (1999): 62-73.

[3]. Rani, Paul Jasmin, Jason Bakthakumar, B. Praveen Kumaar, U. Praveen Kumaar, and Santhosh Kumar. "Voice controlled home automation system using natural language processing (NLP) and internet of things (IoT)." In *2017 Third International Conference on Science Technology Engineering & Management (ICONSTEM)*, pp. 368-373. IEEE, 2017.

[4]. Qiu, Xipeng, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. "Pre-trained models for natural language processing: A survey." *Science China Technological Sciences* 63, no. 10 (2020): 1872-1897.

[5]. Mihalcea, Rada, Hugo Liu, and Henry Lieberman. "NLP (natural language processing) for NLP (natural language programming)." In *Computational Linguistics and Intelligent Text Processing: 7th International Conference, CICLing 2006, Mexico City, Mexico, February 19-25, 2006. Proceedings 7*, pp. 319-330. Springer Berlin Heidelberg, 2006.

[6]. Yandell, Mark D., and William H. Majoros. "Genomics and natural language processing." *Nature Reviews Genetics* 3, no. 8 (2002): 601-610.

[7]. Nadkarni, Prakash M., Lucila Ohno-Machado, and Wendy W. Chapman. "Natural language processing: an introduction." *Journal of the American Medical Informatics Association* 18, no. 5 (2011): 544-551.

[8]. Hirschberg, Julia, and Christopher D. Manning. "Advances in natural language processing." *Science* 349, no. 6245 (2015): 261-266.

[9]. Jones, Karen Sparck. "Natural language processing: a historical review." *Current issues in computational linguistics: in honour of Don Walker* (1994): 3-16.

[10]. Ruder, Sebastian, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. "Transfer learning in natural language processing." In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*, pp. 15-18. 2019.

[11]. Kang, Yue, Zhao Cai, Chee-Wee Tan, Qian Huang, and Hefu Liu. "Natural language processing (NLP) in management research: A literature review." *Journal of Management Analytics* 7, no. 2 (2020): 139-172.

[12]. Yin, Wenpeng, Katharina Kann, Mo Yu, and Hinrich Schütze. "Comparative study of CNN and RNN for natural language processing." *arXiv preprint arXiv:1702.01923* (2017).

[13]. Egger, Roman, and Enes Gokce. "Natural Language Processing (NLP): An Introduction: Making Sense of Textual Data." In *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*, pp. 307-334. Cham: Springer International Publishing, 2022.

[14]. Qiu, Xipeng, Qi Zhang, and Xuan-Jing Huang. "Fudannlp: A toolkit for chinese natural language processing." In *Proceedings of the 51st annual meeting of the association for computational linguistics: system demonstrations*, pp. 49-54. 2013.

[15]. Ly, Antoine, Benno Uthayasooriyar, and Tingting Wang. "A survey on natural language processing (nlp) and applications in insurance." *arXiv preprint arXiv:2010.00462* (2020).

[16]. Jones, Karen Sparck. "Natural language processing: a historical review." *Current issues in computational linguistics: in honour of Don Walker* (1994): 3-16.

[17]. Friedman, Carol, and George Hripcsak. "Natural language processing and its future in medicine." *Acad Med* 74, no. 8 (1999): 890-5.

[18]. Nazir, Farhana, Wasi Haider Butt, Muhammad Waseem Anwar, and Muazzam A. Khan Khattak. "The applications of natural language processing (NLP) for software requirement engineering-a systematic literature review." *Information Science and Applications 2017: ICISA 2017 8* (2017): 485-493.

[19]. Cook, Benjamin L., Ana M. Progovac, Pei Chen, Brian Mullin, Sherry Hou, and Enrique Baca-Garcia. "Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid." *Computational and mathematical methods in medicine* 2016 (2016).

[20]. Ly, Thomas, Carol Pamer, Oanh Dang, Sonja Brajovic, Shahrukh Haider, Taxiarchis Botsis, David Milward, Andrew Winter, Susan Lu, and Robert Ball. "Evaluation of Natural Language Processing (NLP) systems to annotate drug product labeling with MedDRA terminology." *Journal of biomedical informatics* 83 (2018): 73-86.

[21]. Khan, Nabeel Sabir, Adnan Abid, and Kamran Abid. "A novel natural language processing (NLP)–based machine translation model for English to Pakistan sign language translation." *Cognitive Computation* 12 (2020): 748-765.

[22]. Hameed, Ibrahim A. "Using natural language processing (NLP) for designing socially intelligent robots." In *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pp. 268-269. IEEE, 2016.

[23]. Yim, Wen-wai, Meliha Yetisgen, William P. Harris, and Sharon W. Kwan. "Natural language processing in oncology: a review." *JAMA oncology* 2, no. 6 (2016): 797-804.

[24]. Sintoris, Konstantinos, and Kostas Vergidis. "Extracting business process models using natural language processing (NLP) techniques." In *2017 IEEE 19th conference on business informatics (CBI)*, vol. 1, pp. 135-139. IEEE, 2017.

[25]. Galassi, Andrea, Marco Lippi, and Paolo Torroni. "Attention in natural language processing." *IEEE transactions on neural networks and learning systems* 32, no. 10 (2020): 4291-4308.

[26]. Friedman, Carol, and Noémie Elhadad. "Natural language processing in health care and biomedicine." *Biomedical Informatics: Computer Applications in Health Care and Biomedicine* (2014): 255-284.

[27]. Velupillai, Sumithra, Hanna Suominen, Maria Liakata, Angus Roberts, Anoop D. Shah, Katherine Morley, David Osborn et al. "Using clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances." *Journal of biomedical informatics* 88 (2018): 11-19.

[28]. Gonzalez-Hernandez, Graciela, Abeed Sarker, Karen O'Connor, and Guergana Savova. "Capturing the patient's perspective: a review of advances in natural language processing of health-related text." *Yearbook of medical informatics* 26, no. 01 (2017): 214-227.

[29]. Popowich, Fred. "Using text mining and natural language processing for health care claims processing." *ACM SIGKDD Explorations Newsletter* 7, no. 1 (2005): 59-66.

[30]. Iroju, Olaronke G., and Janet O. Olaleke. "A systematic review of natural language processing in healthcare." *International Journal of Information Technology and Computer Science* 8 (2015): 44-50.

[31]. Liu, Feifan, Chunhua Weng, and Hong Yu. "Natural language processing, electronic health records, and clinical research." *Clinical Research Informatics* (2012): 293-310.

[32]. Althoff, Tim, Kevin Clark, and Jure Leskovec. "Large-scale analysis of counseling conversations: An application of natural language processing to mental health." *Transactions of the Association for Computational Linguistics* 4 (2016): 463-476.

[33]. Stewart, Robert, and Sumithra Velupillai. "Applied natural language processing in mental health big data." *Neuropsychopharmacology* 46, no. 1 (2021): 252.

[34]. Koleck, Theresa A., Caitlin Dreisbach, Philip E. Bourne, and Suzanne Bakken. "Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review." *Journal of the American Medical Informatics Association* 26, no. 4 (2019): 364-379.

[35]. Toyabe, Shin-ichi. "Detecting inpatient falls by using natural language processing of electronic medical records." *BMC health services research* 12, no. 1 (2012): 1-8.

[36]. Imler, Timothy D., Justin Morea, and Thomas F. Imperiale. "Clinical decision support with natural language processing facilitates determination of colonoscopy surveillance intervals." *Clinical Gastroenterology and Hepatology* 12, no. 7 (2014): 1130-1136.

[37]. Kang, Yue, Zhao Cai, Chee-Wee Tan, Qian Huang, and Hefu Liu. "Natural language processing (NLP) in management research: A literature review." *Journal of Management Analytics* 7, no. 2 (2020): 139-172.

[38]. Crossley, Scott A., Laura K. Allen, Kristopher Kyle, and Danielle S. McNamara. "Analyzing discourse processing using a simple natural language processing tool." *Discourse Processes* 51, no. 5-6 (2014): 511-534.

