International Journal of Multidisciplinary Research And Studies

Available online at https://ijmras.com/

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH AND STUDIES
ISSN: 2640 7272
Volume:04; Issue:05 (2021)

NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING

Prakash Singh

M.Phil., Roll No. :150127: Session: 2015-16

University Department of COMPUTER SCIENCE, B.R.A. Bihar University, Muzaffarpur, India.

E-mail-: hi.prakashsingh@gmail.com

INTRODUCTION

Language is a means of communication that enables us to read, write, and even converse verbally with one another. For instance, we think, make judgements, and form plans, as well as a variety of other things, in natural language; more specifically, in words. However, the most important problem that we face in this age of AI is determining whether or not it is possible for humans and computers to interact in a way that is comparable. In other words, is it possible for humans and computers to interact with one another using the human language? The development of natural language processing systems is a problem for us since computers need organized data, yet human speech is unstructured and often confusing in its nature. In this sense, we might say that Natural Language Processing (NLP) is the sub-field of Computer Science, in particular Artificial Intelligence (AI), that is concerned with allowing computers to interpret and process human language. AI is the more general term for this area of research. Programming computers to be capable of analyzing and processing vast amounts of natural language data is the primary objective of NLP from a purely technical stand point.

STUDY OF HUMAN LANGUAGES

Language is a crucial component for human lives and also the most fundamental aspect of our behavior. We can experience it in mainly two forms – written and spoken. In the written

 Prakash singh *, *University Department of COMPUTER SCIENCE, B.R.A. Bihar University, Muzaffarpur, India. E-mail: hi.prakas.singh@gmail.com.*

form, it is a way to pass our knowledge from one generation to the next. In the spoken form, it is the primary medium for human beings to coordinate with each other in their day-to-day behavior. Language is studied in various academic disciplines. Each discipline comes with its own set of problems and a set of solution to address those

AMBIGUITY AND UNCERTAINTY IN LANGUAGE

The term "ambiguity," which is often employed in the context of natural language processing, may be defined as the capacity to be comprehended in more than one manner. To put it another way, we may argue that ambiguity is the capacity to be interpreted in more than one manner. This definition is rather straightforward. There is a lot of room for interpretation in natural language. The following categories of ambiguity may be found in NLP:

Lexical Ambiguity: The uncertainty that may be created by only one word is referred to as lexical ambiguity. Consider the use of the word "silver" in the following contexts: noun, adjective, and verb.

Syntactic Ambiguity: The numerous ways in which a statement might be processed can give rise to this kind of ambiguity. Consider the following statement as an illustration: "The guy noticed the girl with the telescope." It is unclear if the guy spotted the girl holding a telescope or whether he saw her via his own telescope when he saw the girl.

Semantic Ambiguity: This kind of ambiguity arises when the actual meaning of the words being used may be taken in a variety of different directions. In other words, semantic ambiguity occurs when a sentence comprises a word or phrase that may be interpreted in several ways. For instance, the line "The vehicle struck the pole while it was moving" has semantic ambiguity because it may be interpreted in two different ways: either "The car hit the pole while it was moving" or "The automobile hit the pole while the pole was moving." Both of these readings are possible".

Anaphoric Ambiguity: This form of ambiguity is brought about as a result of the use of anaphora elements inside conversation. Take, for instance, the horse that galloped up the hill. It was really steep. It became weary very quickly. In this case, the ambiguity is caused by the anaphoric reference to "it" in two different scenarios.

Prakash singh *, *University Department of COMPUTER SCIENCE, B.R.A. Bihar University, Muzaffarpur, India. E-mail: hi.prakas.singh@gmail.com.*

Pragmatic ambiguity: The term "such type of ambiguity" is used to describe the circumstance in which the surrounding context of a word provides it more than one meaning. When a statement lacks specificity, there may be room for pragmatic ambiguity. This may be explained more simply by saying that. For instance, the phrase "I like you too" might be construed in a variety of ways, including "I like you just as much as you like me," "I like you in the same way that you like me," and so on (just like someone else dose).

NLP PHASES

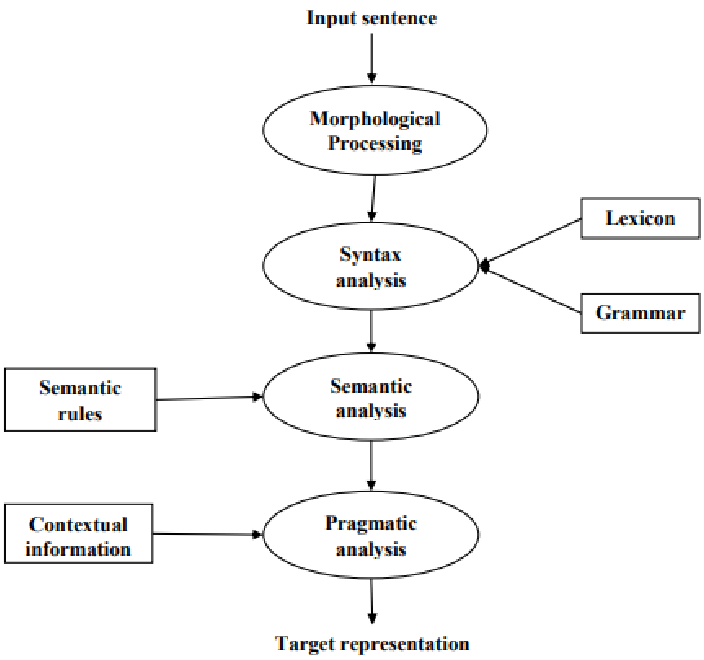

Figure 1. diagram shows the phases or logical steps in natural language processing:

MACHINE LEARNING

This machine learning lesson will cover both the fundamentals and more advanced topics of machine learning. Our instructional guide for machine learning is written with both students and working professionals in mind.

Machine learning is a rapidly developing field of computer science that gives computers the ability to teach themselves new skills by analyzing historical data. The process of developing mathematical models and generating predictions based on past data or information is called machine learning, and it makes use of a variety of methods. At the moment, it is being used for a variety of activities, including picture identification, voice recognition, email filtering, auto-tagging on Facebook, a recommender system, and many more applications.

This machine learning lesson will provide you with an introduction to machine learning as well as a broad variety of machine learning methods including supervised learning, unsupervised

Prakash singh *, *University Department of COMPUTER SCIENCE, B.R.A. Bihar University, Muzaffarpur, India. E-mail: hi.prakas.singh@gmail.com.*

learning, and reinforcement learning. You will acquire knowledge about regression and classification models, as well as clustering approaches, hidden Markov models, and a variety of sequential models.

Need for Machine Learning

The use of machine learning is becoming ever more important. The necessity for machine learning arises from the fact that this kind of technology is able to complete activities that are much too complicated for a human to carry out directly on their own. Because we are just human, we are constrained in certain ways, such as the fact that we cannot manually access the vast amounts of data. Because of this, we need some computer systems, and here is where machine learning comes in to make things simpler for us.

We are able to train machine learning algorithms by presenting them with a vast quantity of data, allowing them to explore the data on their own, building models on their own, and automatically predicting the output that is needed. The effectiveness of the algorithm for machine learning is dependent on the quantity of data, which may be established via the use of the cost function. We will be able to save both time and money with the assistance of machine learning.

RESEARCH METHOLODOGY

The sort of machine learning technology known as supervised learning is perhaps the one that is utilized in natural language processing applications the most frequently. The fact that the learner has access to the values of the target function f for a specific number of instances facilitates the inductive process that is used in supervised learning, which is the method by which an approximation known as f is constructed. This collection of cases is known as the training set, and the values associated with each of them are often determined through the process of human annotation. In order to decide whether or not supervised learning is an effective choice of machine learning technique for an application, it is necessary to consider whether or not it is possible to gather a big enough number of examples that are representative of the target function. During this lecture, we will examine a number of supervised machine learning techniques and demonstrate them by describing in detail how they might be used to the classification of text.

Prakash singh *, *University Department of COMPUTER SCIENCE, B.R.A. Bihar University, Muzaffarpur, India. E-mail: hi.prakas.singh@gmail.com.*

DATA ANALYSIS

UNSUPERVISED LEARNING AND APPLICATIONS

In the second lecture, we discussed several categorization activities that may be learned through supervised practice. Learning was accomplished by the use of a training set, in which the labelling of cases determined the goal function, and an approximation of this function was carried out by the classifier. The unlabeled instance served as the input, while a classification judgement was sent as the output. Unsupervised learning, in which the idea of "learning" is interpreted in a somewhat different way, is presented in this lecture. The learning process makes an effort, as it has in the past, to generate a generalization; but, this time around, no explicit approximation of a goal function is built1. Finding natural groups within a data collection is the key to achieving the generalization that you are looking for. The analysis of unlabeled examples is the only method that can lead to the discovery of these categories. Exploratory data analysis, also known as data mining, has been known to make use of unsupervised learning. In this type of analysis, clustering may be used to uncover patterns of association in the data, providing a foundation for the visualization of association patterns. Examples of such information visualization approaches include dendrograms and self-organizing maps, to name just a couple. In addition, unsupervised learning has been utilized in a few subfields of natural language processing, such as information retrieval, object and character recognition, and dimensionality reduction for the purpose of text classification.

DATA REPRESENTATION

It is essential to have a consistent data representation model while engaging in unsupervised learning, just as it is when engaging in supervised learning. It is possible to utilize the generic vector space model that was covered in lecture 2 in this context as well. As was the case previously, the makeup of the feature set as well as the manner in which values are allocated are unique to the learning job. Clustering of documents, for example, might be done on the basis of real-valued feature vectors constructed in the manner outlined in section 2.2.4. The feature set might be the set of kinds, and the full data set may be interpreted as a co-occurrence matrix if word clustering were to be based on vectors of word co-occurrence in documents.

$$
\begin{array}{lccccccc}
 & & \text{lecture} & \text{we} & \text{examined} & \text{clustering} & \text{groups} & \ldots \\
\text{lecture} & = \langle & 2, & 2, & 1, & 2, & 0 & ,\ldots \rangle \\
\text{we} & = \langle & 2, & 2, & 1, & 2, & 0 & ,\ldots \rangle \\
\text{examined} & = \langle & 1, & 1, & 1, & 2, & 0 & ,\ldots \rangle \\
\text{clustering} & = \langle & 2, & 2, & 1, & 3, & 1 & ,\ldots \rangle \\
\text{groups} & = \langle & 0, & 0, & 0, & 1, & 1 & ,\ldots \rangle \\
\vdots & & & & \vdots & & &
\end{array}
$$

Prakash singh *, *University Department of COMPUTER SCIENCE, B.R.A. Bihar University, Muzaffarpur, India. E-mail: hi.prakas.singh@gmail.com.*

Figure 2 : Co-occurrence vector representation for words

If we took, for example, and this section up to this sentence as our document set, adopting a co-occurrence vector representation for words would give us a representation along the lines shown in Figure 2. This would be the case if we took, and this section up to this sentence as being our document set. One's first instinct may lead them to believe that the word "groups" is peculiar in some way, and that the word "clustering" is the term that most closely resembles it in meaning. The fundamental goal of unsupervised learning is to identify some kind of link like this.

MAIN ALGORITHMS

The most important category of unsupervised learning strategies is represented by clustering algorithms. Hierarchical clustering and partitional clustering are the two primary categories that Jain et al. (1999) use to categories them as a whole. Different hierarchical clustering algorithms have different approaches to calculating the distances that separate groups while the techniques are being used in the clustering process. The hierarchical clustering techniques that are employed the most frequently are the single-link, complete-link, and average-link approaches. Clustering techniques such as mode searching, graph theoretic clustering, and k-means clustering are all examples of partitioned clustering. Another example is the Expectation Maximization method. A distinction that is orthogonal to the one drawn between agglomerative and divisive clustering can also be made in terms of the cluster creation technique that is used (Jain et al., 1999). Agglomerative methods are used to build hierarchical structures from the ground up, beginning with each instance being placed in its own unique cluster and gradually constructing a structure by the use of consecutive merging procedures. On the other hand, divisive clustering places all of the instances into a single group to begin with, and then constructs the structure by dividing the group into smaller and smaller clusters until a predetermined stopping point is reached. When you group data instances together, you have to evaluate how "near" the instances are to one another. In other words, it entails determining the distances that separate two different occurrences. In hierarchical clustering, the similarity between groups is used as the criterion for combining smaller groups of examples into larger groups. This similarity is estimated as a function of the distances that separate pairs of instances, one of which comes from each group.

Algorithm 1: Simple agglomerative hierarchical clustering

Prakash singh *, *University Department of COMPUTER SCIENCE, B.R.A. Bihar University, Muzaffarpur, India. E-mail: hi.prakas.singh@gmail.com.*

```
hclust(D: set of instances): tree
  var: C,  /* set of clusters */
       M  /* matrix containing distances between */
          /* pairs of clusters                   */
  for each d ∈ D} do
     make d a leaf node in C
  done
  for each pair a,b ∈ C do
     M_{a,b} ← d(a,b)
  done
  while (not all instances in one cluster) do
    Find the most similar pair of clusters in M
    Merge these two clusters into one cluster.
    Update M to reflect the merge operation.
  done
  return C
```

Distance and dissimilarity measures

A concept of "distance" between the items in the data set is implied by clustering, as was discussed in the section that came before this one. This idea will be formalized farther down. Given that examples a, b, and c are all represented as real-valued vectors, as was just explained, we define the distance between instances a and b to be a function that satisfies the following conditions and calls itself d(a, b):

$$\begin{aligned} d(a,b) &\geq 0 \\ d(a,a) &= 0 \\ d(a,b) &= d(b,a) \\ d(a,b) &\leq d(a,c)+d(b,c) \end{aligned}$$

If the requirement (3.4), sometimes known as the "triangular inequality," is not met, then the value d is considered to be dissimilar. In the case of agglomerative approaches, clustering algorithms make use of these measurements in order to group (in the case of clustering) objects (instances and clusters) together inside a tree structure.

Hierarchical clustering

The result of an algorithm for hierarchical clustering is a tree structure known as a dendrogram. In this structure, similarity between clusters is represented by linkages between (sister) nodes, and the height of the links indicates the level of similarity between the clusters. A straightforward and all-purpose method for agglomerative clustering is outlined in Algorithm 1.

Prakash singh *, *University Department of COMPUTER SCIENCE, B.R.A. Bihar University, Muzaffarpur, India. E-mail: hi.prakas.singh@gmail.com.*

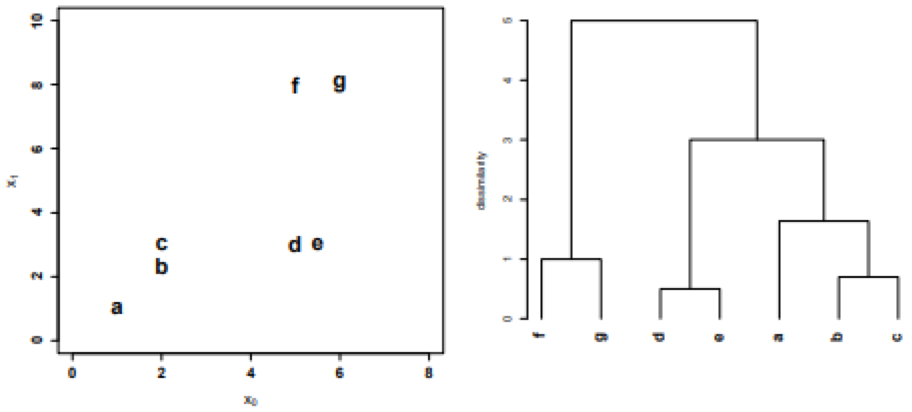

Figure 3 Clustering in 2-dimensional space

CONCLUSION

Research on learning machines has been conducted from a wide variety of perspectives, often under a plethora of different titles. It is not unheard of for a technique or approach associated with machine learning to have been investigated in the past under a different name at some point in time. For instance, the fields of artificial intelligence (AI) and statistics are two examples of these types of fields. In the field of artificial intelligence (AI), the phrase "machine learning" was used in the late 1970s to refer to a collection of methods that were developed to automate the process of acquiring new knowledge. These methods were used to teach computers new information. Although some machine learning techniques, such as neural nets, have been around since the 1950s, the term "machine learning" as it is used today wasn't first used to designate these techniques until the late 1970s. Some machine learning techniques, such as neural nets, have been around since the 1950s. Consolidation of the area was assisted forward by theoretical developments in computational learning theory and the revival of connectionism in the 1980s. As a consequence of these advancements, the field, which drew its motivation and ideas from a wide variety of sources, including information theory, statistics, and probabilistic inference, was able to expand and evolve.

REFERENCES

1. Abend, O. and A. Rappoport (2017). The state of the art in semantic representation. See acl (2017).
2. Abney, S., R. E. Schapire, and Y. Singer (1999). Boosting applied to tagging and PP attachment. See emn (1999), pp. 132–134.
3. Abney, S. P. (1987). The English noun phrase in its sentential aspect. Ph. D. thesis, Massachusetts Institute of Technology.

 Prakash singh *, *University Department of COMPUTER SCIENCE, B.R.A. Bihar University, Muzaffarpur, India. E-mail: hi.prakas.singh@gmail.com.*

4. Abney, S. P. and M. Johnson (1991). Memory requirements and local ambiguities of parsing strategies. Journal of Psycholinguistic Research 20(3), 233–250.
5. Adafre, S. F. and M. De Rijke (2006). Finding similar sentences across multiple languages in wikipedia. In Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources.
6. Ahn, D. (2006). The stages of event extraction. In Proceedings of the Workshop on Annotating and Reasoning about Time and Events, pp. 1–8. Association for Computational Linguistics.
7. Aho, A. V., M. S. Lam, R. Sethi, and J. D. Ullman (2006). Compilers: Principles, Techniques, & Tools (2nd ed.). Addison-Wesley Publishing Company.
8. Aikhenvald, A. Y. (2004). Evidentiality. Oxford University Press.
9. Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6), 716–723.
10. Akmajian, A., R. A. Demers, A. K. Farmer, and R. M. Harnish (2010). Linguistics: An introduction to language and communication (Sixth ed.). Cambridge, MA: MIT press.
11. Alfano, M., D. Hovy, M. Mitchell, and M. Strube (2018). Proceedings of the second acl workshop on ethics in natural language processing. In Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing. Association for Computational Linguistics.
12. Alfau, F. (1999). Chromos. Dalkey Archive Press.

Prakash singh *, *University Department of COMPUTER SCIENCE, B.R.A. Bihar University, Muzaffarpur, India. E-mail: hi.prakas.singh@gmail.com.*