A review on ambiguity of words in NLP

Sandeep Kaur, Dr. Rakesh Kumar Principal, Prabhjeet Kaur Assistant Professor

Computer Science Engg. ,PTU Jalandhar

Sandeepbhutta25@gmail.com

Prabhgill59@gmail.com

Abstract— Building a computer system that can understand human languages has been one of the long-standing goals of artificial intelligence. Currently, most state-of-the-art natural language processing (NLP) systems use statistical machine learning methods to extract linguistic knowledge from large, annotated corpora. However, constructing such corpora can be expensive and time-consuming due to the expertise it requires to annotate such data. In this thesis, we explore alternative ways of learning which do not rely on direct human supervision. In particular, we draw our inspirations from the fact that humans are able to learn language through exposure to linguistic inputs in the context of a rich, relevant, perceptual environment. In this paper NLP interface, various techniques for synonym are described.

Keywords—

I. INTRODUCTION

Building a computer system that can understand human languages has been one of the long-standing goals of artificial intelligence. Currently, most state-of-the-art natural language processing (NLP) systems use statistical machine learning methods to extract linguistic knowledge from large, annotated corpora. However, constructing such corpora can be expensive and time-consuming due to the expertise it requires to annotate such data. In this thesis, we explore alternative ways of learning which do not rely on direct human supervision. In particular, we draw our inspirations from the fact that humans are able to learn language through exposure to linguistic inputs in the context of a rich, relevant, perceptual environment. We first present a system that learned to sportscast for RoboCup simulation games by observing how humans commentate a game. Using the simple assumption that people generally talk about events that have just occurred, we pair each textual comment with a set of events that it could be referring to. By applying an EM-like algorithm, the system simultaneously learns a grounded language model and aligns each description to the corresponding event. The system does not use any prior language knowledge and was able to learn to sportscast in both English and Korean. Human evaluations of the generated commentaries indicate they are of reasonable quality and in some cases even on par with those produced by humans.

For the sportscasting task, while each comment could be aligned to one of several events, the level of ambiguity was low enough that we could enumerate all the possible alignments. However, it is not always possible to restrict the set of possible alignments to such limited numbers. Thus, we present another system that allows each sentence to be aligned to one of exponentially many connected subgraphs without explicitly enumerating them. The system first learns a lexicon and uses it to prune the nodes in the graph that are unrelated to the words in the sentence. By only observing how humans follow navigation instructions, the system was able to infer the corresponding hidden navigation plans and parse previously unseen instructions in new environments for both English and Chinese data. With the rise in popularity of crowdsourcing, we also present results on collecting additional training data using Amazon's Mechanical Turk.

Since our system only needs supervision in the form of language being used in relevant contexts, it is easy for virtually anyone to contribute to the training data. Being able to communicate with a computer in human languages is one of the ultimate goals of artificial intelligence (AI) research. Instead of learning special commands or control sequences (e.g. a series of mouse clicks, typing, or gestures), we could articulate what we want in our own words. In response, the computer could also present information to us or ask questions verbally without those responses having been programmed into the system. In order to achieve this goal, there are two tasks the computer must become competent at: the ability to interpret human languages and the ability to generate coherent natural language content.

II. RELTED WORK

Varun Chandola, Eric Eilertson, Levent ErtAoz, GyAorgy Simon, Vipin Kumar[1] presented that data mining brings a set of tools and techniques that can be applied to discover hidden patterns that provide healthcare professionals an additional source of knowledge for making decisions. In more detail, clustering the patients that have the same status helps discovering new disease, but the suitable number of clusters is not often obvious. This paper first reviews existing methods for selecting the number of clusters for the algorithm. Then, an improved algorithm is presented for learning k while clustering.

Mahesh Kumar Kond Reddy, Sujeeth .T[2] defined that it is a major issue to retrieve good websites from the larger collections of websites. As the number of available web pages grows, it is become more difficult for users finding documents relevant to their interests. Clustering is the classification of a data set into subsets (clusters), so that the data in each subset share some common trait -often proximity according to some defined distance measure. By clustering we improve the quality of websites by grouping similar websites in groups. This paper addresses the applications of data mining tool Weka byapplying k means clustering to find clusters from huge data sets and find the attributes that govern optimization of search engines. Unlabeled document collections are becoming increasingly common and mining such databases becomes a major challenge

Hamzeh Agahi, A. Mohammadpour, S. Mansour Vaezpour[3] Grouping data into meaningful clusters is very important in data mining. K-means clustering is a fast method for finding clusters in data. The integral inequalities are a predictive tool in data mining and k-means clustering. Many papers have been published on speeding up k-means or nearest neighbor search using inequalities that are specific for Euclidean distance. An extended inequality related to Hölder type for universal integral is obtained in a rather general form.

Ahmed Elgohary, Ahmed K. Farahat, Mohamed S. Kamel, and Fakhri Karray[4] provided an overview about k-mean that kernel k-means is an effective method for data clustering which extends the commonly-used k-means algorithm to work on a similarity matrix over complex data structures. The kernel k-means algorithm is however computationally very complex as it requires the complete data matrix to be calculated and stored. Further, the kernelized nature of the kernel k-means algorithm hinders the parallelization of its computations on modern infrastructures for distributed computing. This paper is defining a family of kernel-based low-dimensional embeddings that allows for scaling kernel k-means on Map Reduce via an efficient and unified parallelization strategy. Afterwards, we propose two methods for low-dimensional embedding that adhere to our definition of the embedding family. Exploiting the proposed parallelization strategy, we present two scalable MapReduce algorithms for kernel k-means..

Chunfei Zhang, Zhiyi Fang[5] depicted that The traditional K-means algorithm is a widely used clustering algorithm, with a wide range of applications. This paper introduces the idea of the K-means clustering algorithm analysis the

advantages and disadvantages of the traditional K-means clustering algorithm elaborates the method of improving the K-means clustering algorithm based on improve the initial focal point and determine the K value. Simulation experiments prove that the improved clustering algorithm is not only more stable in clustering process, at the same time, improved clustering algorithm to reduce ore even avoid the impact of the noise data in the dataset object to ensure that the final clustering result is more accurate and effective.

Christopher Ndehedehe, Ogunlade Simeon, Akwaowo Ekpa[6] defined that data mining is the application of specific algorithms for extracting patterns from data. Different Data mining techniques have been used on large volumes of data to discover hidden patterns and relationships helpful in decision making. This work investigates the reliability of K-means, a popular and simplest unsupervised learning algorithm in Land Use Land Cover mapping of Uyo Capital City. The spatial subset of the classified imagery and the ground truth data sampled for this work was a 500m x 500m window. K-means classification was done using different iterations for the five clusters identified in the study area. The confusion matrix, overall accuracy and kappa coefficient results were good. The overall accuracies were 95.835% and 97.588% while the kappa coefficients were0.95 and 0.97 for 50 and 80 iterations respectively. The results were also confirmed by overlaying the various cluster groups with other validated data sources like Ortho photo and digitized vector of the same location. The use of K-means clustering analysis in land use classification may provide us with significant findings and reliable classification results like the supervised and machine learning algorithms.

M.Sakthi, Antony Selvadoss Thanamani[7] presented that due to the increase in the quantity of data across the world, it turns out to be very complex task for analyzing those data. Categorize those data into remarkable collection is one of the common forms of understanding and learning. This leads to the requirement for better data mining technique. These facilities are provided by a standard data mining technique called Clustering. The key intention of this technique is to categorize a dataset into a set of clusters that contains similar data items, as computed by some distance function. One of the widely used clustering techniques is K-Means clustering. K-Means clustering is very simple and effective for clustering. But, the main disadvantage of this technique is when the large dataset is used for clustering. To overcome this difficulty, various researchers focus on suggesting better alteration in K-Means clustering. This paper provides a new technique to modify K-Means clustering which can result in better performance. For initialization, this paper uses an improved version of Hopfield Artificial Neural Network (HANN) algorithm. Also, the Genetic Algorithm (GA) is in combined with K-Means algorithm.

Muhammad Rukunuddin Ghalib, Shivam Vohra, Sunish Vohra, Akash Juneja[8] In data mining, classification is a form of data analysis that can be used to extract models describing important data classes. Two of the known learning algorithms used are Naïve Bayesian (NB) and SMO (Self-Minimal-Optimisation) .Thus the following two learning algorithms are used on a Car review database and thus a model is hence created which predicts the characteristic of a review comment after getting trained. It was found that model successfully predicted correctly about the review comments after getting trained. Also two clustering algorithms: K-Means and Self Organising Maps (SOM) are used and worked upon a Car Database (which contains the properties of many different CARS), and thus the following two results are then compared.

Asmita Yadav[9] Data mining is the process of taking out of concealed prognostic information from a huge amount of databases. It is an influential technology which helps companies to focus on important information in their data warehouses. There are different steps in data mining process like Anomaly detection, Association rule learning, Clustering, Classification, Regression,

Summarization. This paper is mainly concerned about clustering which is the procedure of organizing the objects in groups whose members contains some kind of similarity. In the present review work, will make an attempt for identifying the major issues and challenges associated with different clustering algorithms.

Soumi Ghosh, Sanjay Kumar Dubey[10] data mining technology has been considered as useful means for identifying patterns and trends of large volume of data. This approach is basically used to extract the unknown pattern from the large set of data for business as well as real time applications. It is a computational intelligence discipline which has emerged as a valuable tool for data analysis, new knowledge discovery and autonomous decision making. The raw, unlabeled data from the large volume of dataset can be classified initially in an unsupervised fashion by using cluster analysis i.e. clustering the assignment of a set of observations into clusters so that observations in the same cluster may be in some sense be treated as similar. The outcome of the clustering process and efficiency of its domain application are generally determined through algorithms. There are various algorithms which are used to solve this problem. In this research work two important clustering algorithms namely centroid based K-Means and representative object based FCM (Fuzzy C-Means) clustering algorithms are compared.

S. No	Author	Title	Contribution
1	Farhan et. al. [12]	TOM: Twitter Opinion Mining Framework using Hybrid Classification Scheme, Decision Support Systems	focused on various primary issues like accuracy, data sparsity and sarcasm problems and presents an algorithm for twitter feeds classification based on a hybrid approach. Then proposed method includes various pre-processing steps before feeding the text to the classifier.
2	G. et. al. [13]	Machine Learning Based Approach to Enhance the Accuracy of Sentiment Analysis	examined how classifiers work while doing opinion mining over Twitter data. Reducing the data size using the feature selection method produces better accuracy and increase the computational space. The feature selection method plays a vital role in increasing the accuracy of sentiment analysis.
3	Pedro et. al. [15]	NILC_USP : A hybrid system for sentiment analysis in Twitter Messages	adopted a hybrid classification process that uses three classification approaches: rulebased, lexicon-based and machine learning approaches. They suggest a pipeline architecture that

			extracts the best characteristics from each classifier.
4	Amit et. al. [16]	Feature Extraction for Sentiment Classification on Twitter Data	introduced a novel approach for automatically classifying the sentiment of "tweets" into positive, negative and neutral sentiment. Experimental evaluations show that proposed techniques are efficient and perform better than previously proposed methods.
5	Shoushan et. al. [17]	Sentiment Classification and Polarity Shifting	examined how classifiers work while doing opinion mining over Twitter data. Reducing the data size using the feature selection method produces better accuracy and increase the computational space.
6	K. Revathy et. al. [18]	A Hybrid Approach for Supervised Twitter Sentiment Classification	presented more significant approach towards the contextual information in the document which is one of the drawbacks of the systems which are available for determining contextual information.
10	Yu-Long Qiao et al. [19]	Improved K Nearest Neighbor Classification Algorithm	proposed a technique to reduce the complexity of K-NN classification by using approximation coefficient of a fully decomposed feature vector with Haar wavelet and the variance of the corresponding untransformed vector, to produce two efficient test conditions.

III. LEVELS OF NLP

The most explanatory method for presenting what actually happens within a Natural Language Processing system is by means of the 'levels of language' approach. This is also referred to as the synchronic model of language and is distinguished from the earlier sequential model, which hypothesizes that the levels of human language processing follow one another in a strictly sequential manner. Psycholinguistic research suggests that language processing is much more dynamic, as the levels can interact in a variety of orders. Introspection reveals that we frequently use information we gain from what is typically

thought of as a higher level of processing to assist in a lower level of analysis. For example, the pragmatic knowledge that the document you are reading is about biology will be used when a particular word that has several possible senses is encountered, and the word will be interpreted as having the biology sense. Of necessity, the following description of levels will be presented sequentially. The key point here is that meaning is conveyed by each and every level of language and that since humans have been shown to use all levels of language to gain understanding, the more capable an NLP system is, the more levels of language it will utilize.

A. Phonology:

This level deals with the interpretation of speech sounds within and across words. There are, in fact, three types of rules used in phonological analysis:

1) Phonetic rules: It is used for sound within words.

2) Phonemic rules: It is used for variations of pronunciation when words are spoken together.

3) Prosodic rules: It is used to check for fluctuation in stress and intonation across a sentence.

In an NLP system that accepts spoken input, the sound waves are analyzed and encoded into a digitized signal for interpretation by various rules or by comparison to the particular language model being utilized.

B. Morphology: Morphology is the first stage of analysis once input has been received. It looks at the ways in which words break down into their components and how that affects their grammatical status. Morphology is mainly useful for identifying the parts of speech in a sentence and words that interact together. The following quote from Forsberg gives a little background on the field of morphology. Morphology is a systematic description of words in a natural language. It describes a set of relations between words' surface forms and lexical forms. A word's surface form is its graphical or spoken form, and the lexical form is an analysis of the word into its lemma (also known as its dictionary form) and its grammatical description. This task is more precisely called inflectional morphology. Being able to identify the part of speech is essential to identifying the grammatical context a word belongs to. In English, regular verbs have a ground form with a limited set of modifications, however, irregular verbs do not follow these modification rules, and greatly increase the complexity of a language. The information gathered at the morphological stage prepares the data for the syntactical stage which looks more directly at the target language's grammatical structure.

1) Syntax: Syntax involves applying the rules of the target language's grammar, its task is to determine the role of each word in a sentence and organize this data into a structure that is more easily manipulated for further analysis. Semantics are the examination of the meaning of words and sentences.

a) Grammar: In English, a statement consists of a noun phrase, a verb phrase, and in some cases, a prepositional phrase. A noun phrase represents a subject that can be summarized or identified by a noun. This phrase may have articles and adjectives and/or an embedded verb phrase as well as the noun itself. A verb phrase represents an action and may include an imbedded noun phrase along with the verb. A prepositional phrase describes a noun or verb in the sentence. The majority of natural languages are made up of a number of parts of speech mainly: verbs, nouns, adjectives, adverbs, conjunctions, pronouns and articles.

b) Parsing: Parsing is the process of converting a sentence into a tree that represents the sentence's syntactic structure. The statement: "The green book is sitting on the desk" consists of the noun phrase: "The green book" and the verb phrase: "is sitting on the desk." The sentence tree would start at the sentence level and break it down into the noun and verb phrase. It would then label the articles, the adjectives and the nouns. Parsing determines whether a sentence is valid in relation to the language's grammar rules.

C. Semantics: It builds up a representation of the objects and actions that a sentence is describing and includes the details provided by adjectives, adverbs and propositions. This process gathers

information vital to the pragmatic analysis in order to determine which meaning was intended by the user. D. Pragmatics: Pragmatics is "the analysis of the real meaning of an utterance in a human language, by disambiguating and contextualizing the utterance". This is accomplished by identifying ambiguities encountered by the system and resolving them using one or more types of disambiguation techniques .

1) Ambiguity: Ambiguity is explained as "the problem that an utterance in a human language can have more than one possible meaning.

Types of Ambiguity: Syntactic Ambiguity is present when more than one parse of a sentence exists. "He lifted the branch with the red leaf." The verb phrase may contain "with the red leaf" as part of the imbedded noun phrase describing the branch or "with the red leaf" may be interpreted as a prepositional phrase describing the action instead of the branch, implying that he used the red leaf to lift the branch.

- Semantic Ambiguity is existent when more than one possible meaning exists for a sentence as in "He lifted the branch with the red leaf." It may mean that the person in question used a red leaf to lift the branch or that he lifted a branch that had a red leaf on it.
- Referential Ambiguity is the result of referring to something without explicitly naming it by using words like "it", 'he" and "they." These words require the target to be looked up and may be impossible to resolve such as in the sentence: "The interface sent the peripheral device data which caused it to break", it could mean the peripheral device, the data, or the interface.
- Local Ambiguity occurs when a part of a sentence is unclear but is resolved when the sentence as a whole is examined. The sentence: "this hall is colder than the room," exemplifies local ambiguity as the phrase: "is colder than" is indefinite until "the room" is defined.

IV. CONCLUSION

Summary form only given. Natural language processing (NLP) is a major area of artificial intelligence research, which in its turn serves as a field of application and interaction of a number of other traditional AI areas. Until recently, the focus in AI applications in NLP was on knowledge representation, logical reasoning, and constraint satisfaction - first applied to semantics and later to the grammar. In the last decade, a dramatic shift in the NLP research has led to the prevalence of very large scale applications of statistical methods, such as machine learning and data mining. Naturally, this also opened the way to the learning and optimization methods that constitute the core of modern AI, most notably genetic algorithms and neural networks. In this paper we give an overview of the current trends in NLP and discuss the possible applications of traditional AI techniques and their combination in this fascinating area.

V. REFERENCES

[1]. Varun Chandola, Eric Eilertson, Levent ErtÄoz, GyÄorgy Simon and Vipin Kumar, "Data Mining for Cyber Security," Data Warehousing and Data Mining Techniques for Computer Security, Springer, 2006

[2]. Mahesh Kumar Kond Reddy, Sujeeth .T, "Data Mining Tool using Clustering Technique on Exploration Engine Dataset", Int. Journal of Engineering Research and Application, ISSN : 2248-9622, Vol. 3, Issue 5, Sep-Oct 2013, pp.2032-2036

[3]. Hamzeh Agahi, A. Mohammadpour, S. Mansour Vaezpour, "Predictive tools in data mining and k-means clustering: Universal Inequalities", June 2013, Volume 63, Issue 3-4, pp 779-803

[4]. Ahmed Elgohary, Ahmed K. Farahat, Mohamed S. Kamel, and Fakhri Karray, "Embed and Conquer: Scalable Embeddings for Kernel k-Means on MapReduce", arXiv:1311.2334v2 [cs.LG] 12 Nov 2013

[5]. Chunfei Zhang, Zhiyi Fang, "An Improved K-means Clustering Algorithm", Journal of Information & Computational Science 10: 1 (2013) 193–199

[6]. Christopher Ndehedehe, Ogunlade Simeon,1Akwaowo Ekpa, "Spatial Image Data Mining Using K-Means Analysis: A Case Study of Uyo Capital City, Nigeria", International Journal of Advanced Research (2013), Volume 1, Issue 7, 6-15, ISSN NO 2320-5407

[7]. M.Sakthi, Antony Selvadoss Thanamani, "An Enhanced K Means Clustering using Improved Hopfield Artificial Neural Network and Genetic Algorithm", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-2, Issue-3, July 2013

[8]. Muhammad Rukunuddin Ghalib, Shivam Vohra, Sunish Vohra, Akash Juneja, "MINING ON CAR DATABASE EMPLOYING LEARNING AND CLUSTERING ALGORITHMS", International Journal of Engineering and Technology (IJET)

[9]. Asmita Yadav, " A Survey Of Issues And Challenges Associated With Clustering Algorithms", International Journal for Science and Emerging, Technologies with Latest Trends, 10(1): 7-11 (2013)

[10]. Soumi Ghosh, Sanjay Kumar Dubey, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms", International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013

[11]. Kavita OZA, Rajanish KAMAT, "Applying Data Mining for Framing of Computer Science Curriculum", Proceedings of the IETEC'13 Conference, Ho Chi Minh City, Vietnam. 2013

[12]. Farhan Hasan Khan "TOM: Twitter Opinion Mining Framework using Hybrid Classification Scheme, Decision Support Systems".

[13]. G. Vaitheeswaran , L. Arockiam " Machine Learning Based Approach to Enhance the Accuracy of Sentiment Analysis " International Journal of Computer Science and Management Studies , Vol. 4 , Issue 5 , 2016.

[14]. Namita Mittal , Basant Agarwal "A Hybrid Approach for Twitter Sentiment Analysis".

[15]. Pedro P. B. Filho ,Thoago A. S. Pardo "NILC_USP : A hybrid system for sentiment analysis in Twitter Messages " Internatonal workshop on Semantic Evaluation,2014.

[16]. Amit G. Shirbhate ,Sachin N. Deshmukh "Feature Extraction for Sentiment Classification on Twitter Data" International Journal of Science and Research.

[17]. Shoushan Li, Sophia Yat Mei Lee, Ying Chen , Chu-Ren Huang , Guodong Zhou" Sentiment Classification and Polarity Shifting".

[18]. K. Revathy, B. Sathiyabhama " A Hybrid Approach for Supervised Twitter Sentiment Classification" International Journal of Computer Science and Business Informatics.

[19]. Dhanashri Chafale, Amit Pimpalkar" Sentiment Analysis on Product Reviews Using Plutchik's Wheel of Emotions with Fuzzy Logic" An International Journal of Engineering & Technology, Vol. 1, No. 2 ,December, 2014.

[20]. M. Govindarajan"Sentiment Analysis of Movie Reviews using Hybrid Method of Naive Bayes and Genetic Algorithm "International Journal of Advanced Computer Research, Vol.3, Issue-13, December-2013