International Journal of Recent Engineering Research and Development (IJRERD)
ISSN: 2455-8761
www.ijrerd.com || Volume 04 – Issue 10 || October 2019 || PP. 29-34

An optimized technique to reduce ambiguity in NLP

Sandeep Kaur
Research Scholar,

Prabhjeet Kaur,
Assistant Professor
Computer Science Engg., Sachdeva Girls College Gharuan, PTU Jalandhar

Abstract: Deep learning methods employ multiple processing layers to learn hierarchical representations of data, and have produced state-of-the-art results in many domains. Recently, a variety of model designs and methods have blossomed in the context of natural language processing (NLP). In this paper, we review significant deep learning related models and methods that have been employed for numerous NLP tasks and provide a walk-through of their evolution. We also summarize, compare and contrast the various models and put forward a detailed understanding of the past, present and future of deep learning in NLP. In the research first of all the meaning of all the words are gathered i.e. corpus is to be prepared. After the translation of all the words to Punjabi GA approach comes into play where it choose the best suited meaning during translation. As the best suited meaning is discovered with the help of GA, sentence formation is done using the grammatical rules of language and then reduce the ambiguity. In the research the execution time is calculated to check out the time efficiency of the proposed algorithm. As the execution time may be defined as the total time taken by the proposed algorithm to translate a sentence into Punjabi in an accurate way.

Keywords: GA, Natural Language Processing, Deep Learning, Translation

I. INTRODUCTION

Natural Language Processing (NLP) is a general term for a wide range of tasks and methods related to automated understanding of human languages. In recent years, the amount of available diverse textual information has been growing rapidly, and specialised computer systems can offer ways of managing, sorting, filtering and processing this data more efficiently. As a larger goal, research in NLP aims to create systems that can also `understand' the meaning behind the text, extract relevant knowledge, organise it into easily accessible formats, and even discover latent or previously unknown information using inference. For example, the field of biomedical research can benefit from various text mining and information extraction techniques, as the number of published papers is increasing exponentially every year, yet it is vital to stay up to date with all the latest advancements. Research in Machine Learning (ML) focuses on the development of algorithms for automatically learning patterns and making predictions based on empirical data, and it offers useful approaches to many NLP problems. Machine learning techniques are commonly divided into three categories:

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.

II. RELATED STUDY

[1] Daniel C. Cavalieri et al. (2015) proposed an exponential interpolation to merge a part-of-speech-based language model and a word based n-gram language model to accomplish word prediction tasks. In order to find a set of mathematical equations to properly describe the language modeling, a model based on partial differential equations is proposed. [11]

[2] Karl Pichottaet. al. (2014) Scripts represent knowledge of stereotypical event sequences that can aid text understanding. Initial statistical methods have been developed to learn probabilistic scripts from raw text corpora; however, they utilize a very impoverished representation of events, consisting of a verb and one dependent argument. Author present a script learning approach that employs events with multiple arguments. Unlike previous work, we model the interactions between multiple entities in a script. Experiments on a large corpus using the task of inferring held-out events (the "narrative cloze evaluation") demonstrate that modeling multi-argument events improves predictive accuracy. [3]

[3] Danqi Chen et al. (2013) introduce a neural tensor network (NTN) model which predicts new relationship entries that can be added to the database. This model can be improved by initializing entity representations with word vectors learned in an unsupervised fashion from text, and when doing this, existing relations can even be queried for entities that were not present in the database. [1]

[4] Will Y. Zou et al. (2013) introduce bilingual word embeddings: semantic embeddings associated across two languages in the context of neural language models. Paper propose a method to learn bilingual embeddings from a large unlabeled corpus, while utilizing MT word alignments to constrain translational equivalence. The new embeddings significantly out-perform baselines in word semantic similarity. A single semantic similarity feature induced with bilingual embeddings adds near half a BLEU point to the results of NIST08 Chinese-English machine translation task. [2]

[5] Stephen Roller et al. (2013) improve a two-dimensional multimodal version of Latent Dirichlet Allocation in various ways. (1) outperform text-only models in two different evaluations, and demonstrate that low-level visual features are directly compatible with the existing model. (2) present a novel way to integrate visual features into the LDA model using unsupervised clusters of images. The clusters are directly interpretable and improve on our evaluation tasks. (3) provide two novel ways to extend the bimodal models to support three or more modalities. We find that the three-, four-, and five-dimensional models significantly outperform models using only one or two modalities, and that non textual modalities each provide separate, disjoint knowledge that cannot be forced into a shared, latent structure.[4]

[6] Sergio Guadarramaet. al. (2013) present a solution that takes a short video clip and outputs a brief sentence that sums up the main activity in the video, such as the actor, the action and its object. Unlike previous work, our approach works on out-of-domain actions: it does not require training videos of the exact activity. If it cannot find an accurate prediction for a pre-trained model, it finds a less specific answer that is also plausible from a pragmatic standpoint. We use semantic hierarchies learned from the data to help to choose an appropriate level of generalization, and priors learned from web-scale natural language corpora to penalize unlikely combinations of actors/actions/objects; we also use a web-scale language model to "fill in" novel verbs, i.e. whenthe verb does not appear in the training set.[5]

[7] Karl Pichottaet. al. (2013) addressed the problem of identifying multiword expressions in a language, focusing on English phrasal verbs. Our polyglot ranking approach integrates frequency statistics from translated corpora in 50 different languages. Our experimental evaluation demonstrates that combining statistical evidence from many parallel corpora using a novel ranking-oriented boosting algorithm produces a comprehensive set of English phrasal verbs, achieving performance comparable to a human-curated set. [6]

[8] Shruti Bhosale et al. (2013) presented an approach for detecting promotional content in Wikipedia. By incorporating stylometric features, including features based on n-gram and PCFG language models, we demonstrate improved accuracy at identifying promotional articles, compared to using only lexical information and meta features.[7]

[9] Dan Garrette et al. (2013) Developing natural language processing tools for low-resource languages often requires creating resources from scratch. While a variety of semi-supervised methods exist for training from incomplete data, there are open questions regarding what types of training data should be used and how much is necessary. We discuss a series of experiments designed to shed light on such questions in the context of part-of-speech tagging.[8]

[10] Niveda Krishnamoorthy et al. (2013) present a holistic data-driven technique that generates natural-language descriptions for videos. We combine the output of state-of the- art object and activity detectors with "real world" knowledge to select the most probable subject-verb-object triplet for describing a video. We show that this knowledge, automatically mined from web-scale text corpora, enhances the triplet selection algorithm by providing it contextual information and leads to a four-fold increase in activity identification. Unlike previous methods, our approach can an notate arbitrary videos without requiring the expensive collection and annotation of a similar training video corpus. [9]

[11] Sindhu Raghavan et al. (2013) consider the problem of learning commonsense knowledge in the form of first-order rules from incomplete and noisy natural-language extractions produced by anoff-the-shelf information extraction (IE) system. Much of the information conveyed in text must be inferred from what is explicitly stated since easily inferable facts are rarely mentioned. The proposed rule learner accounts for this phenomenon by learning rules in which the body of the rule contains relations that are usually explicitly stated, while the head employs a less-frequently mentioned relation that is easily inferred. [10]

III. EXISTING SCHEME

Existing schemes for word based prediction developed a natural exponential interpolation model, which combines a traditional word-based n-gram language model with a POS-based language model, defined as the linear combination of three different POS-based languages (with each weight coefficient based on the AUC). It addressed this problem by first finding a partial differential equation to represent the language modeling, which will be used to derive the interpolation model.

It also focused on combining a word n-gram and a m-POS based language model, it is worth noting that there is a growing body of work using continuous-space models in a variety of language processing tasks, particularly for deriving semantic representations of words.

A large portion of the substance accessible in advanced configuration is in English language. The substance appeared in English must be exhibited in a language which can be comprehended by the target group. There is vast area of populace at both national and state level who can't appreciate English language. It has achieved language hindrance in the side lines of computerized age. Machine Translation (MT), can defeat this hindrance. In this postulation, a proposed Statistical Based Machine Translation framework for making an interpretation of English content to Punjabi language has been proposed. English is the source language and the Punjabi is the objective language.

IV. WORD BASED PREDICTION AND POS TAGGING

In word prediction, a statistical language model tries to predict the next word based on the history of previous words.

This idea of word prediction is formalized by probabilistic models called n-gram models, which in turn predict the next word from the $n - 1$ previous words.

In its simplest version, the unigram model only considers the absolute frequency of the word. When using this model, at each moment the most frequent words that begin with written letters of the word in progress are predicted.

part-of-speech tagging (POS tagging or PoS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context—i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph.

A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.

Part-of-speech tagging is harder than just having a list of words and their parts of speech, because some words can represent more than one part of speech at different times, and because some parts of speech are complex or unspoken.

V. GENETIC ALGORITHM

Genetic Algorithm (GA) is a metaheuristic inspired by the process of natural selection that belongs to the larger class of evolutionary algorithms (EA).

Genetic algorithms are commonly used to generate high-quality solutions to optimization and search problems by relying on bio-inspired operators such as mutation, crossover and selection.

In a genetic algorithm, a population of candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem is evolved toward better solutions.

Each candidate solution has a set of properties (its chromosomes or genotype) which can be mutated and altered; traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible.

A typical genetic algorithm requires:

- a genetic representation of the solution domain,
- a fitness function to evaluate the solution domain.

Initialization: The population size depends on the nature of the problem, but typically contains several hundreds or thousands of possible solutions. Often, the initial population is generated randomly, allowing the entire range of possible solutions (the search space). Occasionally, the solutions may be "seeded" in areas where optimal solutions are likely to be found.

Selection: During each successive generation, a portion of the existing population is selected to breed a new generation. Individual solutions are selected through a fitness-based process, where fitter solutions (as measured by a fitness function) are typically more likely to be selected.

Fitness: The fitness function is defined over the genetic representation and measures the quality of the represented solution. The fitness function is always problem dependent.

VI. PROPOSED WORK

In this we described how to adapt discriminative re ranking to improve the performance of the generative models for grounded language learning. Specifically, we delve into the problem of navigational instruction following discussed in last chapter and aid two PCFG models described earlier with the framework of discriminative reranking. Conventional methods of discriminative reranking require gold-standard references in order to evaluate candidates and update the model parameters in the training phase of reranking. However, grounded language learning problems do not have gold-standard references naturally available; therefore, direct application of conventional reranking approaches do not work. Instead, we show how the weak supervision of response feedback (e.g., successful task completion in the navigational task) can be used as an alternative, experimentally demonstrating that its performance is comparable and even more effective compared to training on gold-standard parse trees. Modified Reranking Algorithm for Grounded Language Learning. In reranking, a baseline generative model is first trained and it generates a set of candidate outputs for each training example.

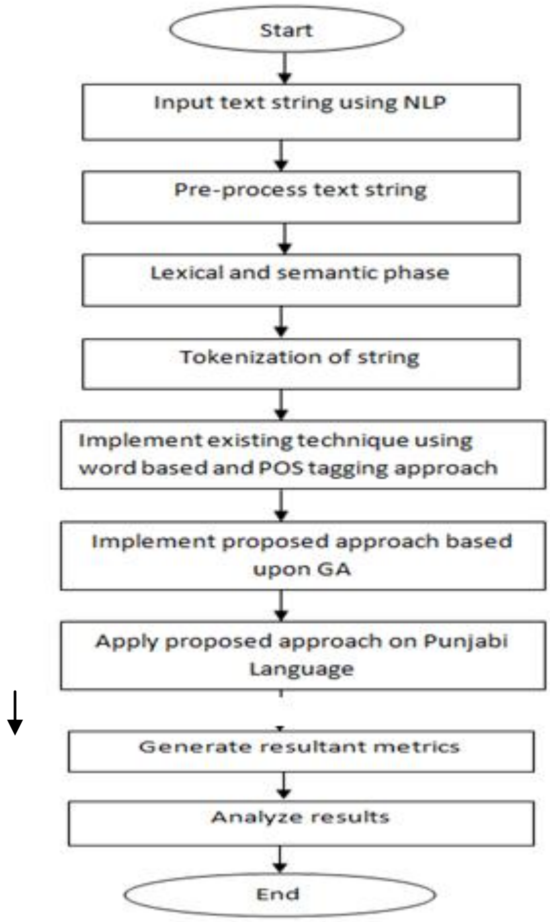

Fig 1: Flow chart

VII. RESULTS

Accuracy: It is a portrayal of orderly mistakes, a measure of factual inclination; as these reason a distinction between an outcome "true" value, ISO calls this trueness.

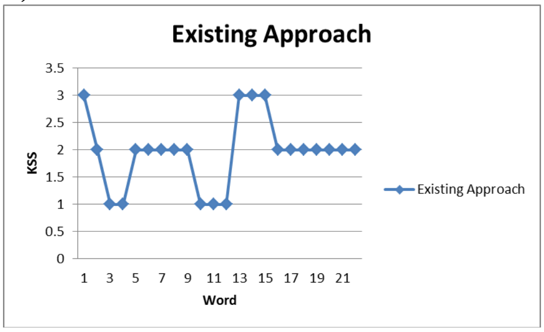

Figure 2: KSS in Existing Approach

Figure 2 is key stroke analysis of the existing approach on 21 number of keywords. Using the existing approach the average value of key stroke is 2

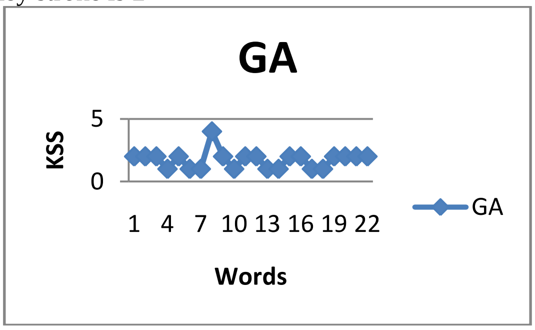

Figure 3: KSS in GA

Figure 3 is key stroke analysis of the existing approach on 21 number of keywords. Using the proposed approach the average value of key stroke is 1.72 which is less than that of existing approach.

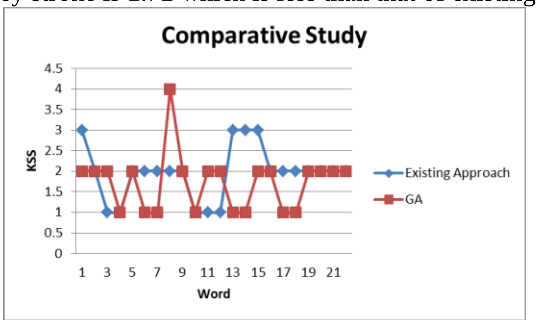

Figure 4: KSS

Figure 4 is key stroke analysis of the existing approach and proposed approach i.e. GA based on 21 number of keywords. Using the proposed approach the average value of key stroke is 1.72 which is less than that of existing approach in which it is approx. 2.

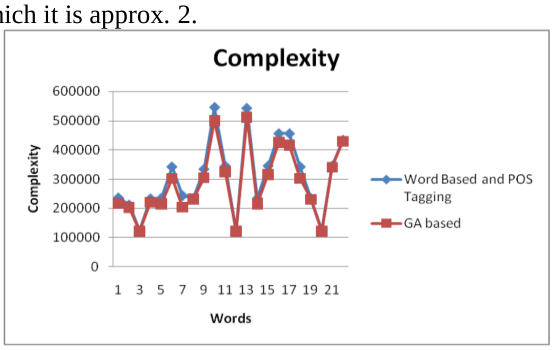

Figure 5: Complexity

Figure 5 is a comparative study of complexity in case of existing and proposed approach. In this figure the complexity in Word based POS tagging that is existing approach is more than that of proposed approach that is GA based approach.

VIII. CONCLUSION

This paper carries out various approaches for network construction in NLP. In this research proposal an optimized technique for prediction analysis has been developed based upon GA.
This approach is efficient in terms of complexity.
Prediction on English language is good.
This technique is also analyzed on a new language Punjabi. It is also efficient on this new language.

IX. REFERENCES

[1]. Danqi Chen, Richard Socher, Christopher D. Manning, Andrew Y. Ng, "Learning New Facts From Knowledge BasesWith Neural Tensor Networks and Semantic Word Vectors", Proceedings of the International Conference on Learning Representations (ICLR, Workshop Track), 16 March 2013

[2]. Will Y. Zou, Richard Socher, Daniel Cer, Christopher D. Manning, "Bilingual Word Embeddings for Phrase-Based Machine Translation", Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)

[3]. Karl Pichotta, Raymond J. Mooney, "Statistical Script Learning with Multi-Argument Events", Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)

[4]. Stephen Roller, Sabine Schulte im Walde, "A Multimodal LDA Model Integrating Textual, Cognitive and Visual Modalities", Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), 1146--1157, Seattle, WA, October 2013.

[5]. Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, "YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-shot Recognition", Proceedings of the 14th International Conference on Computer Vision (ICCV-2013), 2712--2719, Sydney, Australia, December 2013

[6]. Karl Pichotta, John DeNero, "Identifying Phrasal Verbs Using Many Bilingual Corpora", Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), 636--646, Seattle, WA, October 2013.

[7]. Shruti Bhosale, Heath Vinicombe, Raymond Mooney, "Detecting Promotional Content in Wikipedia", Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), 1851--1857, Seattle, WA, October 2013.

[8]. Dan Garrette, Jason Mielens, Jason Baldridge, "Real-World Semi-Supervised Learning of POS-Taggers for Low-Resource Languages", Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL-2013), 583--592, Sofia, Bulgaria, August 2013.

[9]. Niveda Krishnamoorthy, Girish Malkarnenkar, Raymond Mooney, Kate Saenko, Sergio Guadarrama, "Generating Natural-Language Video Descriptions Using Text-Mined Knowledge", Proceedings of the NAACL HLT Workshop on Vision and Language (WVL '13), 10--19, Atlanta, Georgia, July 2013

[10]. Sindhu Raghavan, Raymond J. Mooney, "Online Inference-Rule Learning from Natural-Language Extractions", Proceedings of the 3rd Statistical Relational AI (StaRAI-13) workshop at AAAI '13, July 2013.

[11]. Daniel C. Cavalieri, Sira E. Palazuelos-Cagigas, Teodiano F. Bastos-Filho, Mario Sarcinelli-Filho, “Combination of Language Models for Word Prediction: An Exponential Approach”, IEEE transactions on audio, speech, and language processing, February 2015

[12]. JyotiKumari, Ashok Kumar Dubey, “A Review Paper on Genetic Algorithm”, International Journal of Advance Research in Computer Science and Management Studies, ISSN: 2321-7782 (Online), Volume 4, Issue 7, July 2016, pp: 122-125