International Journal of Advance and Applied Research (IJAAR)
Peer Reviewed Bi-Monthly

ISSN – 2347-7075
Impact Factor – 4.935
Vol.6 No.2 Nov–Dec 2018

N Gram Models for Big Data Development in Machine Learning

Mohd. Khalid Mubashir Uz Zafar[1] Tuba Sabahat[2]

[1]Professor, Department of Translation, Maulana Azad National Urdu University, Hyderabad

[2]Research Scholar, Department of Translation, Maulana Azad National Urdu University, Hyderabad

Abstract - Algorithms are used in machine learning to translate documents written in another language into the user's native language. Machine translations are the result of machine learning, and this machine translation system has a lot of different features and ways to use them. A corpus of words and phrases needs to be developed and tested on various probabilities of algorithm in order to minimize errors in translation for the end user. This study is focused on development of corpus through N-gram modeling, selection of testing method for language model, and the role of big data in machine translations. Information professionals are widely used the advantages of machine translation for satisfying the needs of their users. Although there are no perfect translation systems, the hybrid method is superior to other options because it combines the benefits of several translation methods and improves accuracy to a greater extent.

Keywords: corpus-based, statistical, computational linguistics, machine translation, hybrid machine translation, and language translation

Introduction

The process of translating a source document into another language without the intervention of a human translator is known as machine translation. Machine translation has been around for decades, despite the fact that this is a concept that is relatively new to the general public.

one of the first companies to create machine translation systems in the late 1960s. It worked with the US Air Force to translate intelligence during the Cold War. The objective was for machines to decipher the material so well that human interpreters could grasp its significance and effectively refine the text. Early machine interpretation motors utilized guideline based techniques, which implied that they depended on standards created from people or word references to perform interpretations. Language technology has come a long way since then. Natural Language Processing, also known as NLP, is a subfield of linguistics, computer science, and artificial intelligence that studies how human and computer language interact with one another. It focuses on computer processing and the analysis of a lot of data from natural language. The goal is a computer that can "understand" documents' subtleties, including their language context.

The process of translating one language into another is the focus of machine translation, which is a subfield of both artificial intelligence (AI) and natural language processing (NLP). The statistical machine translation method is widely used in research due to its high accuracy.

Text is translated automatically by computer software in machine translation, eliminating the need for a human translator. The machine translation system and the human translator have had a relationship in which the human translator provided the machine translation system with translations that it had learned from. The speed of the translation process is unquestionably machine translation's greatest strength. The desired translation is available to users within minutes.

Ngram models are broadly utilized in measurable regular phonetics. The N-gram division is used to pattern phonemes—the smallest unit of a word—and their sequences in language recognition. Word patterns are such that there are n words in each gram to analyze. One of the most cutting-edge methods currently utilized in machine translation is this one. The ubiquity of this approach can be checked from the way that the site https://www.academia.edu/alone has over 7.5 million examination papers on NGRM. in connection with research on translation. In Urdu, very little has been done on this front.

N Gram Language Model

The probability of a word appearing at the beginning of a sentence, the word order, or a previous set of words is assigned by linguistic models. Speech recognition, spelling correction, and machine translation are just a few of the applications that make use of linguistic models.

An Overview of the Ngram Linguistic Model: In probabilistic and computational linguistics, n. a set of things. Phonemes, letters, or words are examples of these items. A n-gram series is referred to as a unigram if it is monosyllabic, a bigram if it has two words, and a trigram if it has three words. A one-word n gram is known as one gram in Urdu, while a two-word word is known as two grams, a three-word word is known as three grams, and a four-word word is known as four grams.

Equations can help you understand it.

The equation P (w | h) is the starting point. w probability or chance of comparison, for instanceP (یہ / پانی اتنا شفاف ہے کہ)

Where,

w=یہ

h=پانی اتنا شفاف ہے کہ

Furthermore, the relativity frequency count approach can be used to estimate the aforementioned potential function by counting the number of instances of "that water is so clear" in a relatively large corpus and then counting how many instances occur after the word "it."

Then you'll see how difficult and time-consuming it is to look through the entire corpus. The N-gram model is based on the idea that a possible function can be found using the chain rule. In this case, you will estimate the probability using just a few background words rather than the entire corpus.

Bigram Technique: Using only the conditional probability of a previous work, the Bigram model estimates the probability of all previous words in comparison to the given word. To put it another way, you estimate it using probability: You are

estimating the following when you use the Bigram model to predict the conditional probability of the following word: P (water/it)

$$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-1})$$

The Markov hypothesis is another name for the idea that a word's likelihood is only determined by the word that came before it.
Markov models are a subset of probabilistic models that assume we can predict the probability of a future unit without looking too far back.
The N-gram model can be created by expanding the Bigram model into the Trigram model.
Noisey input can cause incorrect speech-to-text conversions in speech recognition. This can be fixed by N-gram models because they are aware of the possibilities. In a similar vein, N-gram models are utilized in machine translation to produce sentences in the target language that are more natural. Typically, N-gram models are at the word level. Steaming, which is the process of separating the root word from the suffix, is another use for it at the character level. We can also classify languages and distinguish between US and UK spellings by looking at N-gram statistics.
A unigram is a model that doesn't take into account the words before it and only looks at how many times a word appears. A bye occurs when a model uses only the previous word to predict the current word. If you look at the last two words, you'll see that this is a trigram model.
Example: It was raining heavily.
Unigram: (one gram)
موسلا دهار بارش ہورہی تھی
We say that we want to find out how likely the phrase "It was raining heavily" is. The probability can be calculated as follows using the Unigram language model.
P("موسلا دهار بارش ہورہی تھی ") = P(موسلادهار)P(بارش)…P(ہورہی)P(تھی)
P ("It was raining heavily") = P (Heavy rain) P (rain) P (It was raining) P (was)
The probability of each word's presence in the corpus is represented by the above equation. The following formula is used to determine any word's probability:
P(wi)=c(wi)/c(w)
Where w i , i is the word ; c (w i) is the number of w i in the corpus , and c (w) is the number of all words.
Bi gram: (two grams)
was raining heavily
the above sentence as an example and the Bigram language model, the probability can be determined as follows:
P ("It was raining heavily") =
P (rain / beginning of the sentence) P (rain / rain) P (happening / rain). P (ending of the sentence / happening)
The following can be read: The probability that the word "torrential rain" occurred is given to the word "rain." Divide the probability of the word "torrential rain" by the probability of the word itself.
The probability that each of the words that came before it will happen is shown as a product in the example above. The following formula, w i / w i -1, can be used to calculate the probability that any word will occur before any other word:

P (w i w i -1) = P (w i -1, w i) P (w i -1)

The trigram: using the preceding sentence as an illustration and the trigram linguistic model, the probability can be calculated as follows:

P (heavy, the beginning of the sentence), P (raining, heavy), P (happening, rain), and P (ending, the end) are all equivalent to "It was raining heavily." The probability that the word "rain" has occurred is given to the word. The probability of using the word "happening" divides the probability of using the word "rain."

The last two words, w i w i -2 and w i -1, are generalizations of the preceding. The probability of any given word can be calculated as follows:

P (w i / w i -2, w i -1) = P (w i -2, w i -1, w i) / P (w i -2, w i -1)

DEVELOPMENT OF N-GRAM

Numerous natural language processing applications, including statistical machine translation and automatic speech recognition, now make use of N-Gram language models. However, research indicates that N-Gram is reasonable as long as n has a high enough value. Even though NGRAM models are the foundation of modern linguistic modeling for statistical machine translation (SMT) and automated word recognition (ASR) systems like natural language processing (NLP), they are also criticized for having different languages. After experimenting with these models in genres, datasets, and applications for more than two decades, it has been determined that improving these models is a difficult goal. The straightforward N gram model has undergone numerous tests. However, very few of these varieties have been able to perform consistently. Smoothing methods like Good-Turing, Whitten-Bell, and Kneser-Ney have been utilized to estimate probabilities among these types. This is because estimates of maximum probability are still unreliable and overestimate the probability of the rare n-grams that are actually observed, even when using the simplest n-gram hypothesis. The remaining words are either insignificant or null. N items in the fields of probability and computational linguistics. Phonemes, letters, or words are examples of these items. A n-gram series is referred to as a unigram if it is monosyllabic, a bigram if it has two words, and a trigram if it has three words. In Urdu, a single word n gram is called one gram, a two-word is two grams, a three word is three grams and a four-word is four grams.

The field of statistical natural language processing makes extensive use of n-gram models. Phonemes and their order in speech recognition are modeled using n-gram distribution. The words are arranged in such a way that each n-gram has n words, making them easy to analyze.

Building Corpus: A parallel corpus is one in which a set of original texts written in L1 are translated into L2, Ln, and so on. The majority of parallel corpora only contain bilingual data. Texts written in two or more languages that are identical in gender, subject, register, etc. are called "competitive corpora." are connected to the parallel corpora in a strong way. be that as it may, share a similar material.

Bilingual or multilingual parallel corpora contain texts written in both languages. They can be one-dimensional (like when a German text is translated from English to German), two-way (like when a German text is translated from English to German), or multiple dimensional (such as EU text translated into German, Spanish, or French, among other languages).

In particular, statistical machine translation relies heavily on the parallel corpus. The parallel corpus contains a bilingual text and its translation. The original text and its translation into another language are organized by rows.

Mona Baker was the first person to use the term "parallel corpus" in the context of machine translation. A collection of texts that have been translated from one language to another and are machine-readable—that is, they can be read by machines—is referred to as this. encoded in code.) This set of texts can be translated into two or more languages. Simply put, the parallel corpus is a text bilingual. Johansson and Huffland have also referred to this parallel corpus as the translation corpus due to its use in translation. Take note of the slight distinction between the parallel and comparative corpuses. In the parallel corpus, the number of sentences in the target language and the number of sentences in the original language are exactly the same; however, the number in the comparative corpus and the text may differ in some way. Wikipedia is an illustration of a comparative corpus.

Parallel corpora are also utilized for the research of linguistic features and their expression in bilinguals. Comparative Study or Contrastive Analysis are two names for this. It is utilized for analyzing the similarities and differences that exist between the two languages. It demonstrates how the script, culture, and rules of the two languages differ.

Additionally, the number of parallel corpus compilations for corpus-based dictionaries is rising. A parallel corpus is used in this approach to translate words from one language into another.

Machine translation, in which large parallel corpses are created to ensure accurate translation, is the most common application for parallel corpus. These equal corpora comprise of a few crores of words which, subsequent to being assembled, go through different activities. Parallel businesses aid translators in identifying translation alternatives and analogies. Additionally, they aid in word repetition, usage, and the recognition of syntactic patterns. Words and phrases that cannot be translated into the target language are made easier for translators by this. This enables the selection of the most accurate translation. It has emerged as a significant and significant machine translation resource.

In particular, the incredible and multimillion-fold increase in the working speed of the computer and its immense storage capacity has made the corpus the field of machine translation. The use of the Dolsani parallel corpus is a promising tool for the future in the field of machine translation. created the preferred tool or source. In order to preserve, scientifically study, and machine-translate their

languages, speakers of all languages around the world are now turning their attention to the creation of insane corpses.

Methods of Language Modelling

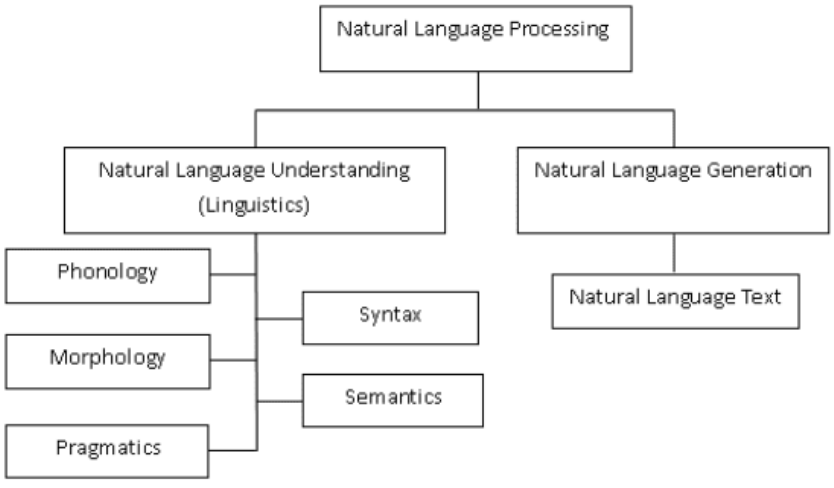
Statistical Language Model: Statistical language modeling or language modeling is the process of creating probabilistic models that can predict the next word in a sequence based on the words that came before it. N-gram language modeling is one example.

Neural Language Model: When it comes to difficult tasks like speech recognition and machine translation, neural network methods are performing better than traditional ones, and this is true for both standalone language models and models that are incorporated into larger models. Using word embeddings, a neural language model can be carried out.

NLP Processing

Natural Language Processing, also known as NLP, is a subfield of linguistics, computer science, and artificial intelligence that studies how human and computer language interact with one another. It focuses on computer processing and the analysis of a lot of data from natural language. A computer that can program to "understand" the nuances of documents' contents, including their language context, is the goal. Using this technology, the documents themselves can be categorized and sorted as well as accurately extracted from the documents.

Word recognition, natural language comprehension, and natural language creation are typically natural language processing challenges. The works that have been studied under the heading of Natural Language Processing are listed below.



Conclusion: The N-Gram model was designed to predict the next word in machine translation. It has been largely successful in this endeavor, but technology has advanced significantly since its conception. In addition to machine translation, neural language models like GPT3 and Bert no longer require a large corpus and have proven to be useful in other natural language works. The Gram model should be replaced with linguistic models based on other neural approaches.

REFERENCES

[1] S. Tripathi, J. K. Sarjgek, "Approaches to machine translation", Annals of Library and Information Studies, Vol. 57, Dec 2010, pp. 388-393.
[2] L.R. Nair, D.S. Peter, P.R. Renjith, "Design and Development of a Malayalam to English Translator-A Transfer Based Approach", IJCL, Vol. 3, Issue. 1, 2012.
[3] C. Dove, O. Loskutova, and R. Fuente, "What's Your Pick: RbMT, SMT or Hybrid?" , 2012, available at: http://amta2012.amtaweb.org/AMTA2012Files/papers/Doveetal.pdf.
[4] F.J. Och, "Minimum Error Rate Training in Statistical Machine Translation", Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, July 2003, pp. 160-167.
[5] M. Nagao, "A Framework Of A Mechanical Translation Between Japanese And English By Analogy Principle". In A. Elithorn and R. Banerji. Artificial and Human Intelligence. Elsevier Science Publishers, 1984.
[6] M. N. Al-Kabi, T. M. Hailat, E.M Al-Shawakfa, I. M Alsmadi, "Evaluating English to Arabic Machine Translation Using BLEU", IJACSA, Vol. 4,2013, pp. 66-73.
[7] E. Sumita, H. Iida, "Experiments and Prospects of Example-based Machine Translation", available at: http://acl.ldc.upenn.edu/P/P91/P91- 1024.pdf.
[8] M. Guidère, "Toward Corpus-Based Machine Translation for Standard Arabic", Translation Journal, vol. 6, no. 1, January 2002.
[9] A.-L Lagarda, V. Alabau, Casacuberta, R. Silva, E. Díaz-de-Liaño, "Statistical Post-Editing of a Rule-Based Machine Translation System". Proceedings of NAACL HLT 2009, pages 217–220.
[10]R.D Brown, "Example Based Machine Translation in Pangloss System", available at: http://www.scism.lsbu.ac.uk/inmandw/ir/example-basedmachine-translation.pdf.
[11]P. Kohen, "Statistical Machine Translation", Cambridge University Press, New York, 2010, pp-53.
[12] S. Nirenburg, H. Somers, Y. Wilks, "Readings in Machine Translation", Asco Typesetters, Hong Kong, pp.157, 233.
[13] T. V. Prasad, G.M. Muthukumaran, "Telugu to English Translation using Direct Machine Translation Approach, International Journal of Science and Engineering Investigations, Vol. 2, Issue. 2, 2013, pp. 25-35.
[14]D. Dinh, N. L. Ngan, D. X. Quang, V. . Nam, "A Hybrid Approach to Word Order Transfer in the English-to-Vietnamese Machine Translation", available at: http://www.amtaweb.org/summit/MTSummit/FinalPapers/58-Dienfinal.pdf.