
Khalid et al LGURJCSIT 2020
ISSN: 2521-0122 (Online)
ISSN: 2519-7991 (Print)
LGU RESEARCH JOURNAL
Computer Sciences IT
Vol.4 Issue3, July-Sep 2020

Domain-Independent Natural Language Processing of text using Unsupervised Translation

Adnan Khalid[1], Aziz Guergachi[2], Qurat ul ain[3], Awais Qasim[1], Adeel Munawar[3]

[1]Department of Computer Science, GC University, Lahore, Pakistan

[2]Ted Roger's School of Management, Ryerson University Toronto, Canada

[3]Department of Computer Science, Lahore Garrison University, Lahore, Pakistan

Abstract

NLP is one of the very important domains of artificial intelligence. Nowadays, advancements are being made and NLP is one of the most developing fields. In this paper, we offer a mutual use of unsupervised translation with n-grams and Natural Language Processing techniques to challenge the difficulty of unsupervised translation extraction from textual data. To build a Text Meaning Extraction System, we have to deliver one important element which is input text. This study presented a different algorithm to work out resemblances between natural languages, by using sequence package analysis and changing them into n-grams. Whenever the sentences that are grammatically difficult and quite lengthy are applied to see the results of the presented algorithm, there are quite efficient results in a semantic reaction. To enhance the experience in the field of AI and search engines, this research paper shows how to improve the handling capability of fuzzy concepts within computers. For example, when search jobs are executed in search engines small textual concepts or sentences might be semantically formed to switch the keyword-based queries. This ability may be functional to intelligent agents to even the procedure of communication between humans and machinery.

Key-words: Text-to-speech; Natural Language Processing; Artificial Intelligence; Text Meaning Extraction System; grammar pattern; information retrieval; Intelligent Agents

1. INTRODUCTION

One of the most widely practiced fields of Artificial Intelligent is Natural Language Processing. There are many applications of natural language processing that have been established in the past few years. Most of the applications of Artificial Intelligence are very good. For example,
a machine that responds to voice and does work according to the instructions given to the machine. There are many researchers doing

research work to develop a new system and applications of AI and make them useful for normal people. There are a lot of tools and different machines available and a lot of work is being done in this field. But the problem with those tools is that they all are domain-specific, there is very little progress in domain-independent natural language processing. NLP is a process or a way to do things like humans do or make machines that can act like humans do and minimize the difference between machines and human beings. Natural Language Processing is a technique where a machine can behave more humanly and thereby reduce the distance between the human being and the machine. Therefore, in simple words, natural language processing helps humans to interact with machines easily.

Now, there are a lot of challenges that we face while processing natural text. Firstly, the term limitations are assorted and then the context of the sentence gets different from its literal meaning. Also, the writing styles and sometimes even spellings of the same words are different in different accents. Sometimes people prefer to write a shorter form of the words such forever is written as 4ever, which is easier for us to understand but a machine is way too dumb to understand its meaning. These are the main challenges faced in natural language processing systems. Programing a model that will have a general understanding of natural language and act accordingly is a very difficult task as we cannot create a data set that will have all the sentences of natural language. Also, there are a lot of problems in natural language. If we consider language ambiguity. Many words have more than 1 meaning, which makes them difficult to understand. These kinds of problems make it very difficult to develop a system like mentions above. The simplest form of the sentence is having two things, a subject and a predicate, predicate is anything which the subject is doing. For English sentences, many of the parts of speech are important to take into consideration such as punctuation, conjunctions, prepositions, adverbs, verbs, nouns, interjections, and at last pronouns. Thus a system should have sufficient knowledge to form a grammatically correct sentence in which the words or phrases are joined correctly and it must have the information of what actual words are. It must also know the literal meanings of the words and how they can be used to form sentences that have different contextual meanings. In addition to all that, the system must know how human beings think and how they reason things. The following are the components that every natural language processing system should know about:

a. Phonological

This deals with the sounds of the words, like how any particular word sounds. It came from the word Phoneme which means the lowest or the smallest entity of sound, these small entities are joined together to form sounds for a complete word.

b. Syntactic

It deals which the structure like how the words are structured in a sentence and what does it mean in that particular forum.

c. World

It is a very important component. This component deals with the surroundings, like to keep going in the conversion without getting cornered. It includes the beliefs of other people and their aims and their sentiments. It also deals

with sentiments like in what sentiment the sentence should be made.

d. Pragmatics

It deals with an advanced level of knowledge acquisition. It deals with the knowledge of how to structure sentences according to the diverse context and how it will affect the meaning of the sentence in other contexts e.g. the word too in the phrases, too hot, and I have different meanings entirely.

e. Morphological

This component deals with verbal understanding, which means the construction of words from their smallest entity known as morphemes. A morpheme is the tiniest entity or meaning. For instance consider the example, the creation of sufficiently from ly and sufficient.

2. LITERATURE REVIEW

Natural language processing has many applications and is rapidly developing. Although a lot of work is being done in this field the main problem of Artificial Intelligence i.e. generality is still not yet solved. After all the research work we did, we were able to find out that Learning Agents and Unsupervised Learning techniques combined can take us a step closer to solving the problem of generality. Before going any further, let's talk about what generality is, and why we need it? Many people criticize Artificial Intelligence as a bag of tricks used in programs to solve ill-defined and difficult problems. That means that every program or system we create can only stick to a set of problems and provide their solutions. However, we need a system that can not only provide solutions to almost all problems but is also able to learn and understand on its own.

To achieve generality we have to work on a few smaller problems. First, we have to work on a small addition to the suite that encompasses the comprehensive revising of the suite with the data structures. Secondly, we still don't have a universal language set of common domain knowledge which could be used for any kind of field and lastly as suggested by John McCarthy that the key problem of achieving generality in artificial intelligence is that we don't have a common language for expressing general common-sense knowledge. Although in the paper that he discussed this problem, a few solutions were also suggested they are not related to text analysis. The pointillism idea for natural language processing was proposed by Peiyou Song and et al. [1] In the proposed model, a very large set of data was divided into the sentences and n-grams and then they were separated into n-grams and sentences. Then those sentences were used to create phrases using exterior evidence for example the presence of n-grams with time-based associations. In another paper written by EfsunSarioglu, Hyeong-Ah Choi, and Kabir Yadav [2] another proposed method/software is MedLee which is mostly used in natural language processing software research community that has truly interpreted results for the raw text procedure-report like CT scan results for brain stroke and pneumonia by chest radiology. the strength of the well-organized output of Medlee is proving codes from UMLS. It is a repository of significantly controlled vocabularies in biomedical developed by the American National Library. Figure1 is showing a sample flow of match keywords mechanism.

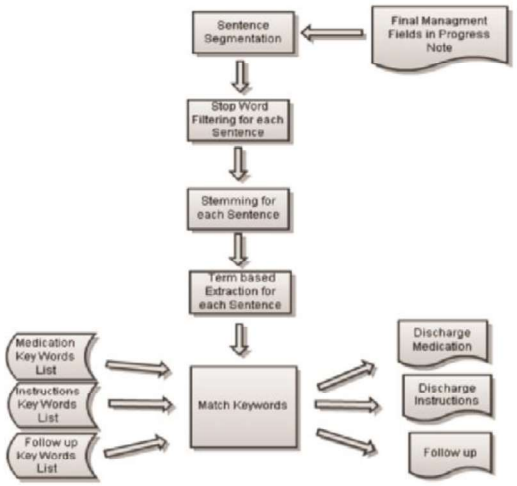

Figure 7: Discharge Medication, Instruction and Follow Up Part

In another study [3] researchers have developed a system that uses semantic comparison calculations and is derived from a unified medical language system, which is called a knowledge-based system. They then evaluated the system. This system assigns a vague concept to other vague terms based on its structure. This process is very important to many other natural language processing systems. We combine the word sense disambiguation methods system with other systems as well. The cTAKES and MetaMap were the two systems that were integrated into the word sense disambiguation system by us. [4]In another paper by Lakshmi K.S and G. Santhosh Kumar, Natural Language Processing tools were combined with data mining algorithms such as the Apriori algorithm and FP-Growth algorithm for the extraction of rules. The extracted rules describe symptoms of a particular disease, an association of the disease with other diseases, medications used for treating diseases, the most prominent age group of patients for developing a particular disease. Another paper [5] presented a semantic analytic technique to spontaneously allocate clinical to complex medical patient data written in the English language. Their semantics assessment method includes dependency parsing of clinical data achieved from training and testing data sets and the calculation of semantic matching scores. Their target was to build an automatic disease identification tool through accurate clinical code assignments ICD-9-CM codes. They evaluated their method with a real-world corpus developed in the 2007 International NLP Challenge administered by the Computational Medicine Center at the Cincinnati Children's Hospital and obtained promising findings.

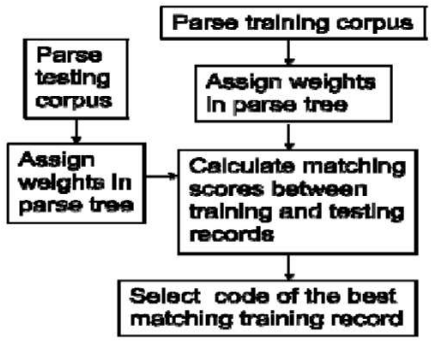

Figure 8: Best Matching Techniques for Parsing [12]

We studied different aspects of Natural Language Processing and how they work in different areas. We mainly studied text meaning extraction with the help of unsupervised learning. We studied in [11] that to form a Text-to-Speech fusion system one must deliver two important mechanisms: an NLP (Natural Language Processing) phase, which functions on the input text, and a language generation phase to yield the anticipated output. These two discrete stages must exchange both data and commands to produce comprehensible and natural language. As the complete TTS job depends on many discrete scientific zones, any accomplishment toward correction can diminish the effort and increase the dynamic of the outcomes.

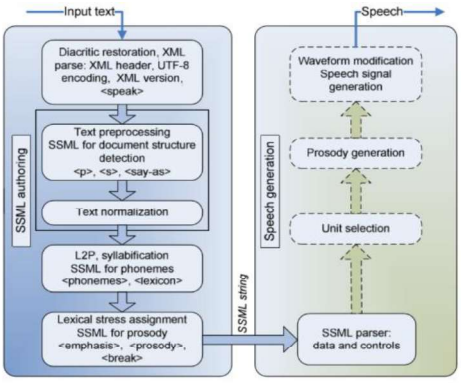

Figure 9: SSML implementation for NLP framework [11]

Then we moved to text encryption in Natural Language Processing. We learned in [12] that the encryption of text is one of the most operational resources for the safety of data. While examining the counterparts of watermarking of text and the encryption of text, an algorithm, which was built based on natural language processing is explained here. Different kinds of language modifications in natural language processing are presented. We learned text distillation in this paper. In this paper, text purifications are explained like how it separates the meaningful data from the cover data. The important steps and methods to extract the textual meaning and purify the text for the bulk is represented in this paper. Meaning extraction of text and watermarking is still not very developed and in its initial phase. We can modify the design of the system of watermarking of natural language to make it work for the extraction of textual meaning. In the end, the necessities and the altered methods for text meaning extraction are delivered. Then at last we moved to our main topic which is the information of writing designs in a specific mass is appreciated in numerous NLP systems like question-answering, organization, information withdrawal, etc. So in this research paper, an uncomplicated actual probabilistic linguistic design for demonstrating in-decomposability of text arrangements under the MDL belief, an efficient unproven knowledge algorithm is executed. The trial on an English critical lettering mass showing auspicious reporting of designs associated with human summary. Some associated investigations may possibly be originated in the training MWE [13] and [14]. The most present examiner focusses on learning MWEs within a partial syntactical arrangement. Possible development deceit in linguistic designs. The present model distinguishes two designs despite even a small alteration such as “We would work” and “We would also work”. Though, many comparable designs are worth bearing in mind not to be so categorized, such as (1) “We would work” and “We would also work”; (2) “in the past” and “in the close past”; (3) “it is possible” and “it is likely”. At the same time, it is not essential to accept to achieve interchangeability for learning non-continuous designs. In detail, many different procedures with interchangeability and by presenting those prototypes, the results may get better and with the help of vocabulary based probabilistic linguistic design, the trick of learning text arrangements are dignified into recognizing, in-decomposability within designs by handling it as an entire term, that fits good into the understanding arrangement of MDL. The linguistic policy is extremely simple in scheming, although it produces honestly cheering results in unremitting form learning. Alteration depiction conserves long-distance reliance in non-continuous designs well. Experimentation results for the non-continuous learning stage seem less prominent than the

constant ones due to the thinness of data, however, the algorithm is still capable to learn many content-independent designs such as “not only ... but (also)”. In a broad-spectrum, the general consequences have wide exposure when likened with a summary from any essay writing textbook. The foremost benefit of this methodology is unverified learning, which entails almost no prior information. Some connected investigation might be discovered in the research of multiword expressions (MWE) [13],[14]. Nevertheless, usually, recent scholars focus on learning MWEs within the partial syntactic configuration such as verb expressions while our method doesn’t take essential structures into attention. Probable development lies in the linguistic model. The modern model distinguishes two patterns despite a minor change such as “We would need” and “We would also need”. Many similar patterns are meaningful, not to be so differentiated, such as “We would need” and “We would also need”; “in the future” and “in the near future”; “it is probable” and “it is expected”. Some random procedures with interchangeability and by presenting those prototypes, the results might get better.

3. METHODOLOGY

We have worked on domain-independent natural language processing. Many tools and systems have given considerable results when applied to a specific domain. Many systems combined with other programs have also given even better results but the problem remains the same i.e. they all are limited to one single domain. Some work is discussed below.

Natural language is understood as a tool that individuals use to express themselves, has exact properties that reduce the efficiency of textual information retrieval systems. These belongings are language discrepancy and vagueness. By language deviation, we mean the likelihood of using unalike words or terminologies to connect the same idea. Dialectaluncertainty is when a word or phrase allows for more than one understanding.

Both occurrences disturb the information recovery process, even though in altered ways. Language difference irritates document muteness, that is, the oversight of relevant brochures that fulfill evidence needs, because the same terms were not used as those found in the document. Vagueness, on the other hand, proposes document noise, or the presence of non-meaningful forms, since documents were recovered that used a similar term but with a dissimilar sense. These features make programmed linguistic handling significantly difficult.

At a morphological level, similar words may show diverse morph-syntactic parts comparative to the situation in which they look, producing uncertainty difficulties, as we can see in example 1mentioned below.

Example 1: A book was the present that his husband gifted her when all of their friends were present at the birthday party.

In this case, the word "present" different meanings and acts both as an adjective and noun. At a grammatical level, concentrating on the study of recognized dealings among words to form larger verbal units, expressions, and verdicts, uncertainties are formed as a result of the option of connecting a sentence with more than one grammatical structure. On the other hand, this variation expects the opportunity of articulating the same idea but altering the direction of the sentence's grammatical structure, as we can see in example 2.

Example 2: She ate the biscuits on the plane.

This example could mean that "She ate the biscuits that were in the plane" or that "She ate the biscuits when she was flying in the plane." Statistical handling of natural language exemplifies the traditional model of information recovery systems and is

categorized from each file's set of keywords, identified as the terms index.

These prototypes are then limited to combining the files' words with that of the request's. Its straightforwardness and efficiency have developed the most frequently used existing models in literal figures recovery schemes. We do not have a system that will take the text from any field and process it according to that field. Every system we have for meaning extraction works for a particular field of a particular domain. It will be efficient to produce a system that will take any language from any field and process is according to that area.

We propose the n-gram model for this problem. N-gram is a statistical language model. In natural language processing, N-gram models are very prominent and are used for statistical analysis of text a lot in this field. It uses a data set to extract the meaning of the text. The two main benefits of *n*-gram models are comparative plainness and the capability to scale up by simply increasing *n,* a model can be used to stock more context with a well-stated space-time tradeoff, enabling small experimentations to measure up very efficiently.

All of this sample exhibit the difficulty of dialects, and that any programmed handling is not easy or observable.

4. EXPERIMENTS AND ANALYSIS

a. Experiments

The main idea of our work is to extract the meaning from the text with the help of the Unsupervised Learning mechanism and learning agents. First, we size the contiguous pairs of sentences and then later on we move to the paragraph without any conclusive result at the sentence level in our experiments.

a. The Unigram

For initial experiments, we applied the n-gram algorithm prototypes to check the word resemblances between two sentences. This is a system in which we only do the handling on one word at a time. The n-gram is a connected system of n items of text. These items can be words, letters, or base pairs. At first, we used unigram which takes sentences and checks if it intersects with any other sentence and we noted that the threshold value for the intersecting is 14%. So when we get results higher than the stated people agree that the sentences have similarities in their topic, but when the intersecting value is close to the 14% the people get confused about their decision.

b. Bigram

A bigram is much like a unigram model but it has value 2 of n. Bigram help provides the possibility of a token. A bigram can be understood as a frame that shows two words at a time.

b. Overlapping of the noun (paragraph level)

At the paragraph level, first, we take out the proper nouns and then both proper and common nouns from each subsection. We can see when the intersection ratio of a proper noun is 10% all the readers recognized the subject as analogous. On the third pair of sections even though the overlay is 50% topic was not recognized specifically and on the other hand, this both proper and common noun extraction presented better outcomes as the overlap percentage of the 3rd pair has gone down to 9.09%. Consequently, we can determine that both common and proper noun extraction gives better performance than only proper noun extraction i.e the noise that happened in the earlier experiment at 1st pair of the sentence has been removed here as the overlap percentage value of 0% from the earlier experiment, has increased up to 16.67%.

We performed several experiments to obtain a better threshold rate and got some refined overlap percentage while extracting both common and proper nouns from each adjacent pair of movements. For the third pair of the

sentence, we get a 20% intersection portion for only proper noun abstraction and most of the human readers were confused about the topic limit. This type of imprecise boundary is not desirable at this greater fraction and after that when we mined both proper and common nouns we have seen that this boundary has gone down to 7.14%. from 20%. So the indistinct topic boundary has come down to a lower percentage level and this is an improvement. However since it is a statistical model, our experiment is not all noise-free.

c. *Experimental methodology*:

We did another experiment for meaning extraction.

a. Preprocessing

Before we do any text extraction work, the document we have to take is a consideration that should have to be processed in a particular form, if we want to get the best results. This process contains steps like tokenization, progression, reducing, and distinguishing discrete chunks of text inside the document, like Title of the document, Headers, Abstract of the document, References, Body of the document, etc. For now, we have done some experiments on Stanford Natural Language Processing suites with some extra rules introduced in it. After extracting the tokens and sentences from the document we built the phrases from them. For example from the sentence "the newly founded government scheme can work efficiently".

Our system to extract tokens from the document mined out 8 key-words from the above-given example, which are: "newly", "newly founded", "newly founded government scheme", "government", "government scheme", “scheme” "work efficiently", “efficiently”. There can be a lot of ways to get a solution after the key-words are mined out. For now, we have selected and applied to simple methods: the first is that in a key-word, two words cannot be separated if there is a verb between them, especially if the verb exist between then, the second is that at the end of each sentence, no commas, full stops, exclamation mark, questions mark or any other can distinct a key-word. This means that the parts of the key-word cannot be from two different parts of a different sentence.

b. Fine-tuning

Stop words that do not have a meaning related to each other cannot be permitted to take part in the process of phrase building. Furthermore, all the symbols of Unicode except some, like minus symbol “-” and semi-colons “ ” cannot be included in the phase of phase building. We only can use the list of stop words of the English language.

c. Reducing

We can dodge the issues like reducing if we eliminate the words which have the same meaning but are in dissimilar forms.

d. Selecting the features set

The utmost and imperative step for extraction procedures is that we choose the framework which satisfies all the needs of the system and works according to the needs of the system. Such a framework is really important. There are enormous features that we can use to perform our task, but we have to select the framework which suits best to our needs

e. Proposed Feature

Each of those words may be labeled as a diverse part of speech by 36 possible labels. Besides part of speech labeling, Stanford's natural language processing system delivers semantic relations among portions of a sentence. So every sentence can be understood as a tree, where words in the sentences are leaves of the tree. The features set we proposed were built to the following recommendations: the first is that we search for key-words in the header of the document, segment headers portions, and reference sections. The previous work was done to limit the complications of problems that could occur and make the system

simpler. If we look at our data, one paper has an average of 2700 words which can be divided furthermore into nearly 8000 phrases. If we take 200 papers for our corpora, we will end up with features that will have more than a million directions or vectors to solve the problem.

f. Result's assessment methodology

The following are the key methods that we use to extract the meaning from textual data: the first of them is Accuracy, the second is Recollection and the third is the G-Measure. If we suppose 3 groups of which we already have mined the meaning. Suppose key-words that are not recognized as key key-words. The appropriately known key-words are represented by the set Y. The mistakenly added key-words are represented by the set Z.

So if we take set X, Y, and Z into consideration. If we call precision U, recall V, and G-Measure G, we can define them as mention here:

$$U = 100\% \ .Y/(Y+Z)$$
$$V = 100\% \ .Y/(Y+X)$$
$$G = 2.UV/(U+V)$$

Now, to regulate and recount the performances of our procedures we will practice the previously explained rules for our system. The important thing to consider is that our calculation was done by comparing the set of already extracted key-words per article with the set, which was chosen by experts for the same document.

Additionally, we measured U, V, and G for just exclusive key-words' existence. For instance, if a known key-words

we consider P, R, F just for exclusive key-phrases occurrence. For instance, if a known key-phrase like "Facebook" comes in a text like 20 times, we will only count it once in our sample set when we are doing the calculations for U, V, and G.

5. CONCLUSION

There are a lot of benefits related to natural language processing applications, the subject is now vastly studied by groups within the technical community. After gesture control systems, NLP is without a doubt the most useful tool we need, as the pace of our lives is getting faster. A lot of progress is being made and many different products are being produced that help in the betterment of everyday life. In the corporate sector, NLP has done a huge part, as there are many people with physical disabilities and it helps ease their life a bit. Also in the health care sector, this commanding tool has assisted in the treatment, surgeries, and recovery of many patients.

One limitation of the present work is in the assumption that somehow we might be able to make a general language, such that we can convert any language or given text into that particular language. To achieve or to make such a language we first need to have a very huge database that can store meanings of all the words in all different languages. To confront this difficult task. first, we need to work on different key-words matching. We should keep in mind that we will have to take consideration of the structure and the meaning of the words if we want to check the key-words which implicitly relate to each other. We assume that the proposed method may also be valid with different and more specific datasets like emails, medical texts, news, different abstracts, web pages etc, as the syntax and or writing styles of people in different fields is entirely different. The validation of this assumption will be our immediate future work.

References

[1] Peiyou Song, AnheiShou, David Phipps, Mohit Tiwari, Dan S. Wallach, Jedidiah R. Crandall, George F. Luger, Language Without Words: A Pointillist Model for NLP.

[2] EfsunSarioglu, Hyeong-Ah Choi and Kabir Yadav, Clicincal Report Classification Using NLP and Topic Modelling, 2012 11th International Conference on Machine Learning and Applications.

[3] Vijay N. Garla and Cynthia Brandt, Knowledge Based Biomedical Word Sense Disambiguation: An Evaluation and Application to Clinical Document Classifaction, 2012 IEEE Second Conference on Healthcare Informatics, Imaging and System Biology.

[4] Lakshmi K.S and G. Santhosh Kumar, Association Rule Extraction from Medical Transcripts of Diabetic Patients.

[5] Ping Chen, Araly Barrera and Chris Rhodes, Semantic Analysis of Free Text and its Application of Automatically Assigning ICD-9-CM Codes to Patient Records, Proc. 9th IEEE International Conference on Cognitive Informatics.

[6] Gerhard WeiB, Learning to Coordinate Actions in Multi-Agent Systems, Institut fur lnformatik, TechnischeUniversitatMiinchenArcisstr. 21, 8000 Miinchen 2, Germany.

[7] James Allen, Nathanael Chambers, George Ferguson, Lucian Galescu, Hyuckchul Jung, Mary Swift, William Taysom, A Collaborative Task Learning Agent, Institute for Human and Machine Cognition, 40 S Alcaniz St, Pensacola, FL, Dept. of Computer Science, Stanford University, Stanford, CA.

[8] Sandra Clara Gadanho, Learning Behavior-Selection by Emotions and Cognition in a Multi-Goal Robot Task, Institute of Systems and Robotics IST, 1049-001 Lisbon, Portugal.

[9] Peter Stone, Learning and Multi-agent Reasoning for Autonomous Agents, Department of Computer Sciences The University of Texas at Austin.

[10]`Geffner, Hector. "Artificial Intelligence: From programs to solvers." AI Communications 27.1 (2014): 45-51.

[11]`Catalin, and DragosBurileanu. "An advanced NLP framework for high-quality Text-to-Speech synthesis." Speech Technology and Human-Computer Dialogue (SpeD), 2011 6th Conference on. IEEE, 2011.