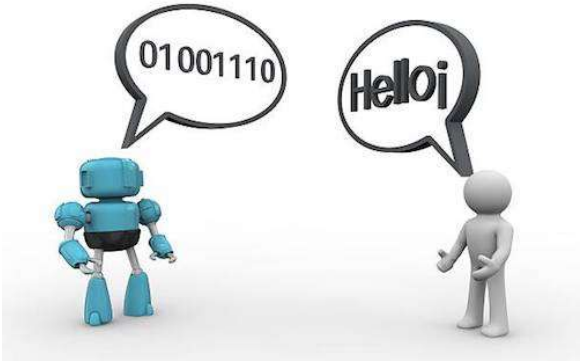
Machines
&
Languages



(An effort to relate languages with machines)

- Sourish Gunesh Dhekane

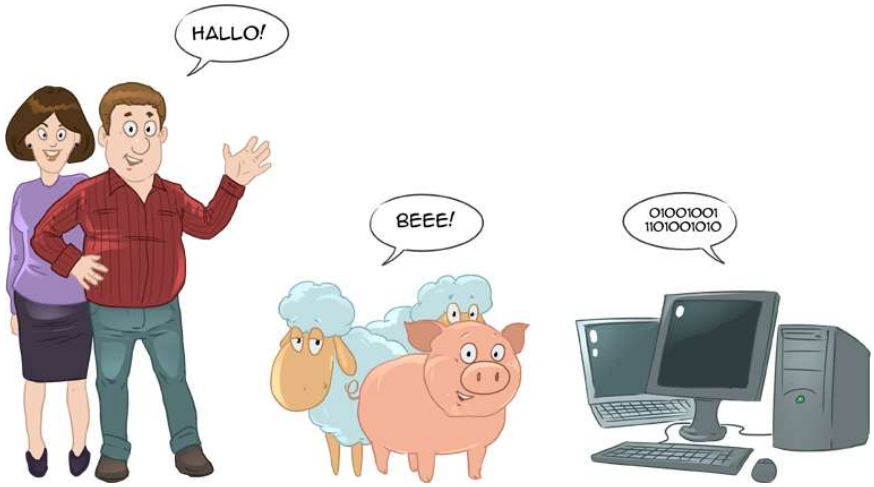
Roll no. 1601020 CSE Dept

Abstract

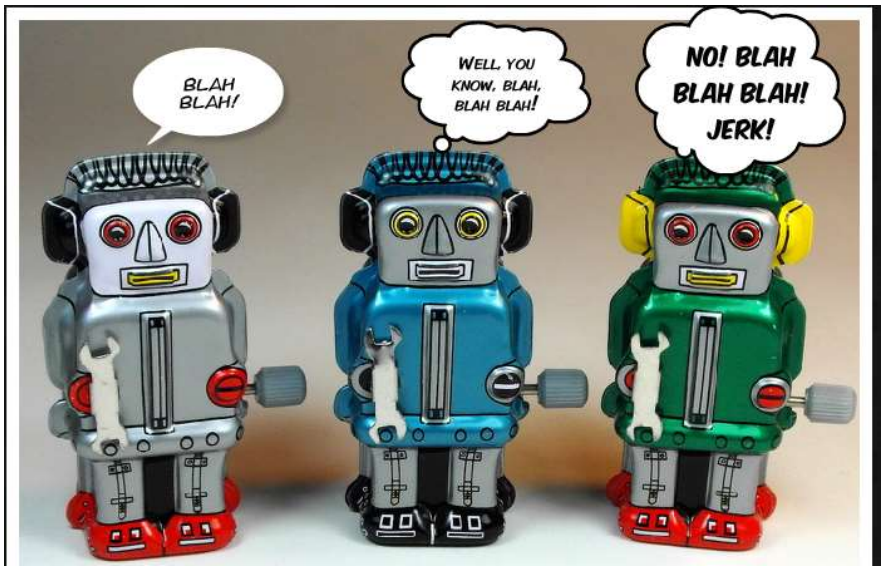
In this paper, we look at the aspects of language associated with machines and modes of communication in them. We also try to relate Language with Intelligent Machines and look at several aspects of Data Driven machines. We then study the field of Natural Language Processing. The main focus of the paper is to put forward two futuristic predictions regarding evolution of language in machines. The views of this paper support the possibility of “Language equipped intelligent machines” in near future. Then we take a look at the consequences of the case where the hypotheses come true, from a linguistic point of view. Lastly, we explore the scope of this field and establish a direction for research in the field of “Machine Linguistics”.

Introduction

Language is an arbitrary, representational and conventionalized system used for HUMAN communication, consisting of linguistic signs which exemplify the relationship between signifier and signified. In this definition of language, all the aspects of language are related to the term HUMAN and thus the definition clearly states language as a characteristic property of humans. Any mode of communication found in any other non-human species is not considered as language. The properties that make humans capable of generating, sustaining and flourishing a language are both biological and intellectual. Lowering of vocal cord or erected spine structure are some of the physical properties that enable humans to generate a variety of sounds and process the information gathered by their senses with greater speed, which are necessary for an accurate expression of emotions for better communication. In addition to this, humans have the ability to think which makes them one step ahead of the rest of the species in terms of communication. Considering the fact that humans have the most developed brains than any other species on the planet, it can be a bit misleading to hypothesize that “Language” is restricted to humans only.



Industrialization has been marked as the most impactful period in human history as human life started becoming more and more efficient due to introduction of MACHINES for doing all the kinds of stuff. As this event can be viewed as a great scientific success, it can also be viewed as the introduction of a “non-living” component (Machines) in the society which in time became a major part of human life. Thus it is an absolute necessity to communicate with machines in order to efficiently lead life. Earlier, the machines were not as complicated as they are now a days. Thus the mode of communication was also much simpler compared to today’s machines.



There is no doubt that communication is present in machines but as stated earlier, there is a difference between mode of communication and language. Mode of communication in machines is something which is predefined (pressing the “ON” button on remote to switch on the AC is a mode of communication but not a language) and predictable (for a specific set of inputs given, machines produce the same output all the time e.g. addition of 5 and 4 on a calculator would give 9 all the time). Considering the above discussion, it would be again very misleading to say that language can’t be associated with machines, as humans have

gone a step ahead in the field of “Artificial Intelligence” and “Machine Learning”.

This paper would take its readers to a journey from the history of machine languages and its aspects to the research done on the intelligent robots for enabling them to have language like us. It would mainly feature some of the important points like the motive of languages in machines, communication in between machines, communication between machines and humans, the concept of programmable feelings and data driven approach and some of the hypothesis or futuristic-predictions. These predictions would be backed by some historical facts and current trends in this topic from a technical point of view. The paper would consist of some technical and non-technical terms of which the readers might not be aware of and thus some useful links for extra reading would be provided where necessary so as to get an idea of the mentioned terms. It is recommended to visit those links for a better understanding of the paper. The rest of the paper is written in an easy language and with analogies to make the tedious technical terms look interesting. This paper would be of immense use for those who are involved in the field of machine learning and new to linguistics. No part of the paper is taken directly as it is from any other source and the thought experiments suggested in this paper are authentic thus plagiarism is discouraged.

Machines

When we try to associate the term Language with machines, we observe that Machine itself is a very complex term and it demands to be defined and explained. Before properly defining the term, we would extract the features of machines that we need for establishing its relation with language and thus the definition would not be a general one but it would definitely serve our purpose.

Firstly, we all know that machines are NON-LIVING things. The meaning of the term "Non-Living" is straight forward i.e. it doesn't contain any "Natural" form of life and it can't generate any natural form of life on its own (Non-Reproducible). This may get confusing if we apply these features to current intelligent bots. These robots are able to produce different machines (thus used in industries). In addition to that, some robots are able to produce copies of their own and therefore one may think that those robots are living. But one should carefully distinguish between the natural form of life and virtual livingness implemented on machines via humans.

See link: - http://news.cornell.edu/stories/2005/05/researchers-build-robot-can-reproduce

If we have a look at the development of machines then we can broadly classify them as MECHANICAL and DIGITAL machines. In this paper, we are less concerned about mechanical machines as they require physical force/ thermal energy/ light waves (forms of mechanical energy) to operate. In digital machines, Signals are the most basic form of communication which are mostly electrical/ electromagnetic in nature. In Digital machines only, we can say that there is a form of communication as there is a specific input pulse and a specific output by the machine. The Digital machines are further classified into NON-

INTELLIGENT machines and INTELLIGENT machines. In Non-Intelligent machines, a specific set of inputs is handled and for a given set of input the answer is always accurate and constant. Calculator is an example of Non-Intelligent machine, as it takes input a specific set of data only (real numbers) and generates an accurate and unique (as there can be only 1 correct answer) output. On the other hand, Intelligent machines are broader in terms of input (they cover a large set of inputs) and produce the most suitable answer according to the ability of the machine (which might not be accurate). They are basically based on the DATA DRIVEN APPROACH (explained later) so that they can learn with each given TRAINING DATA SET. Training data set is nothing but the set of inputs given to the intelligent machines so that they can extract some meaningful features from the data. Our approach would be to correlate Non-Intelligent machines with just a "Mode of Communication" and intelligent machines with "Language".

See link:-

http://encyclopedia2.thefreedictionary.com/intelligent+machine

Therefore for the purpose of our paper, Machines can be defined as NON-LIVING, DIGITAL and HUMAN-MADE objects designed for completing human tasks efficiently.

Why Machines?

This section basically explains the need of “Modes of Communication” and “Language” in Machines. Before the introduction of machines in human lives, “Human Communication” was a necessity. If we look at the origin of language or the motive behind the birth of language we can clearly see that LANGUAGE WAS A NECESSITY FOR SURVIVAL. For the satisfaction of basic needs of humans i.e. Food, Shelter and Cloth, communication is necessary and for that Language is necessary.

Industrialization is seen as the era of introduction of machines in human lives. During initial periods of industrialization, machines were introduced in large scale industries only and thus they were constrained to a small percentage of population only. As people saw potential in machines to serve human needs if manufactured in large quantities, people started inventing machines for completing every task in their day-to-day life. Therefore a large section of population was involved in interactions with machines and MACHINES BECAME PART OF HUMAN LIFE.

Today, humans are totally dependent on machines for almost every need of them and thus it has become a NECESSITY TO LEARN TO COMMUNICATE WITH MACHINES. The above discussion is all about the necessity of “Mode of Communication” in machines. The paper mainly concentrates on “Language” aspect of “Intelligent” machines. To make the point more clear, let us take an example of a cab driver driving a car. If he wants to find the closest gas-station for refilling the vehicle, he can ask his chat bot (Siri/Cortana) to find it just by talking to the device. This is a very simple and somewhat not-so-useful example, but currently many industrial giants like Honda, Intel and NASA are using such intelligent machines for manufacturing purposes. Thus

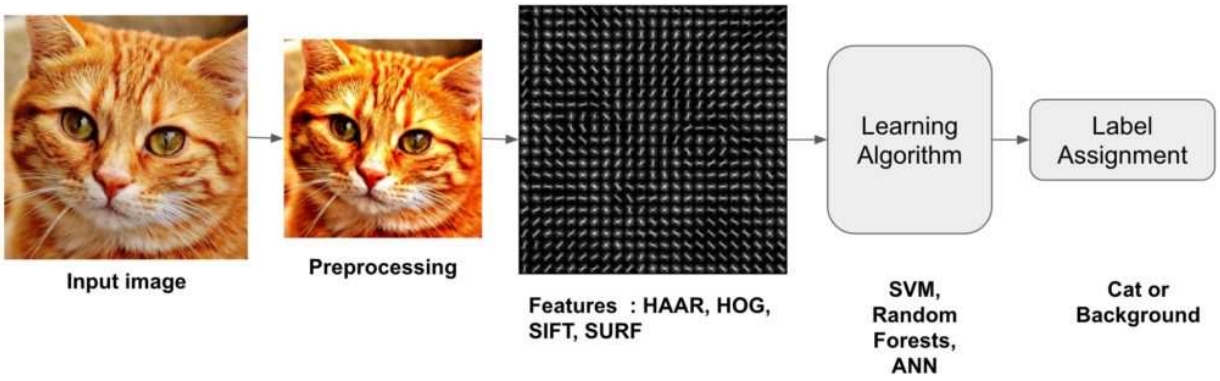
communication between them and with humans also is an absolute necessity.

Let us now know a little about the “Intelligent” machines and their properties so that it becomes much easier to state the hypotheses, which is the motive of this paper.

Intelligent Bots, Machine learning & Data Driven Approach

Machine Learning is a field where machines (algorithms) are designed in such a way that after every iteration of inputs, the algorithm (and thus the machine) learns something new, for which they are not explicitly programmed for. In other words, from the set of given inputs, the algorithm finds the key observations on its own which are known as FEATURES and in technical terms, this process is known as FEATURE EXTRACTION.

See link: - https://en.wikipedia.org/wiki/Feature_extraction.



The output of the machine in the first few attempts is never accurate. The main advantage of this kind of machines is that the output gets better and better with each attempt and thus the machine learns. To get accurate outputs, the algorithm must be trained on a TRAINING DATA.

In machine learning also there are two options i.e. FEATURE IDENTIFICATION and DATA DRIVEN APPROACH. In Feature Identification, the model searches for some key features or properties and it then assigns weights to the deciding properties. On the other hand, the data driven approach doesn't search for any specific feature. Instead it learns directly from the data and searches features on its own. Thus the model is also Data Driven (Driven by the nature of data). In real –life models, the Data Driven model is used as it resembles with the learning process of humans. Consider the example of an infant learning natural language. The development of language in that infant is largely dependent on the data given as an input to him/her (depends largely on the language of his/her parents). This is also an example of Data Driven model of learning.

INTELLIGENT BOTS are the machines designed for certain tasks having Data Driven algorithms. One of the biggest examples of Intelligent Bots is the "Chess playing Device" made by IBM. The model was further developed using CNN (Convolutional Neural Networks) and after enough training data, it was able to defeat the former grandmaster of chess 'Gary Kasparov'.

See link: - https://en.wikipedia.org/wiki/Deep_Blue_(chess_computer)

Today, for several purposes like Object Identification, Future Image Prediction, and Stock Market Predictions etc. these Intelligent Systems are being used. To make the readers understand the magnitude of effect these Intelligent Systems cause on our lives, we take a real world example. “YouTube” is one of the largest video streaming websites and people from every corner of the world watch and upload videos on YouTube. The main reason behind YouTube’s massive success is that the service is free and it is accessible to all i.e. anyone can upload and watch videos on YouTube. This freedom is currently being used for wrong purposes, especially by some terrorist organizations like ISIS which upload offensive videos so provoke and disturb the society. Hundreds of followers of ISIS upload thousands of offensive videos on YouTube and though they get banned and their videos get removed, making another account on YouTube and uploading the videos once again is piece of cake. To manually deactivate such offensive accounts and remove their videos is next to impossible and would not serve the purpose as these videos must not get viral. To deal with this problem, YouTube has come with an Intelligent System which detects offensive videos based on the image content and audio content of the video. They have also massively parallelized such algorithms so that their throughput increases. Today, if any offensive video gets uploaded on YouTube, it is removed and account is permanently banned within a fixed time frame. Imagine what would have happened if these Intelligent Systems were not present! Thus one must accept the importance of such machines.

See link: - https://thenextweb.com/offers/2017/09/27/want-five-years-of-netflix-free-well-we-wanna-give-it-to-you/#.tnw_a8xoBUlB

As we have seen, these Intelligent Systems are designed for a specific purpose. But the progress in this field has gone beyond this point also. Imagine a human like Intelligent System capable of thinking in any given

situation. Interaction with such system would not be of the format “Input-Output”. It would be descriptive and expressive in nature. These systems possess thinking abilities which are not programmed in them by birth (by the time of creation). Such systems are just given the ABILITY TO LEARN and nothing else. Thus after their creation (birth) they behave just like an infant and wait for inputs from several sources to learn different things. Their ability to think depends majorly on the experienced gained by them. Such systems possess an advantage over humans in term of memory and processing speed as memory is accurate and fast (unlike humans) and ability to process gathered data is also way faster than humans. In addition to that, research is being done on inclusion of more and more expressive things in them like Emotions. An illusion of emotions/ thoughts/ ideologies is created in them by something called Behavioral Implementation. Thus we can say that such Robots are a better version of humans and as humans need Language, such systems also need Language for almost the same reasons.

Natural Language Processing

When a newborn tries to communicate with the outside world, the language spoken is unknown to the infant and the relationship between the signifier and signified is vague. While observing the world, the baby CONCIOUSLY tries to find out the relationships between signifier and signified for different linguistic signs. Other than that also, there are some things which are recorded UNCONCIOUSLY like the accent, style of construction of sentences, pauses taken while speaking etc. which influence a lot. Thus there can be several language inputs which may carry the same meaning associated with them.

Natural Language Processing is a field where a large language data consisting of almost all possible styles and combinations is processed by a program (Data-Driven) and efforts are made to interpret it.

See link: - https://en.wikipedia.org/wiki/Natural_language_processing

One of the biggest successful example of it is the “Siri” Chat-Bot. Chat-Bot is an intelligent system which takes input in a specific language which is auditory in nature, tries to find out the meaning associated with the input and then replies to the user in his/her own language.

One of the biggest success of “Siri” is that it covers almost all the accents and works fine for grammatically incorrect inputs also as it should. Imagine a case of a highway accident where the driver is severely injured and not in the position to speak well. Thus an input of “Call an ambulance immediately. My position is mile-marker 47 near Holiday-inn hotel, a mile ahead of gas station” is not expected from the user. A simple set of words like “Accident-Ambulance-Mile Marker 47-Gas Station-Please” should be enough for the chat-bot to understand what has happened. Intelligent Robots (mentioned in the previous section) have a very strong data driven NLP (Natural Language Processing) algorithm which enables them to take literally communicate with humans. Such robots can quickly and permanently learn different languages.

Up to this point, we have covered the language aspects of different types of machines. In the coming sections we move on to the most important point of the paper where we are going to propose some thought experiments and predict their outcomes based on historical facts and current trends. It should be noted that almost no work has been done (people may have worked but no official publications are present till this date) on these topics thus only hypothesis are made in this paper and explanation behind why these are proposed is included. In no way the paper gives formal proofs or assurance that the hypothesis would become correct in distant future. It just provides logical statements depicting a large probability of happening certain events.

Hypothesis 1

Even today, most of the intelligent systems work on the “Input-Output” system i.e. an input is necessary to activate the system. The output might be descriptive in nature but input is necessary. Even if we look at the latest work done in some really amazing fields of AI, the work is restricted in terms of the necessity of an input. We have self-driving cars which take input as a set if images and output a procedure of safe driving. We have GANs (Generative Adversary Networks) which generate an image by taking an input of a paragraph (If we write a paragraph describing a beautiful sunrise the neural net would output an artificially generated image). We have chat bots like Siri/Cortana which can communicate with humans to some extent when provided with a question. But all these cutting edge technologies require some kind of input to activate. This is the major drawback of machines from of point of view of Language.

If we look at the evolution of Language in humans, we find out that language was originated due to the instinct of survival in humans. But as humans socialized and started living in a community, language flourished. Language is not stagnant. Language gets transformed with time. If we take the example of Sanskrit language, it originated hundreds of years ago and after its origination it got shaped and modified as time passed. Many great scholars like *Panini* and *Bhandarkar* tried to establish order and formulate grammatical rules. In addition to such technical modifications, thousands of literary works were made in the forms of *Subhashitas* and *Amarsaaras* mainly. In every language, we can find people who contributed in development and beautification of language through Literature. In Hindi, we have *Premchand* as we have *William Shakespeare* in English. A very

interesting question arises from this discussion. If we give language to an Intelligent Robot, is it possible that the Robot is capable of generation of Literature? Can there be any *Shakespeare* in Robots? We state our first hypothesis in context to this question.

INTELLIGENT DATA DRIVEN ROBOTS ARE CAPABLE OF LITERATURE GENERATION AND IN TIME THE OUTPUT-ORIENTED APPROACH OF SUCH SYSTEMS WILL BE REDUCED.

Here, we state that such Robots are CAPABLE of literature generation though to this date not much progress has been made in this direction. The Output-Oriented approach of such machines means the tendency to complete a task all the time. If we think from the point of view of necessity then a basic form of language is all that is needed in a society for Survival. This statement in no way ignores the importance of literary works in human social life but it also states that for survival, basic form of language should suffice. Thus without any motivation (extreme need), generation of literation happens. Today's intelligent systems lack this property. Still we state a bold statement about their progress. The reason behind making such risky statement can be understood from the coming discussion.

Most of the literary works are based on HUMAN EMOTIONS and Historical Events. Making such systems able to produce literature based on historical events or even events happening in present is very easy and much work has been done in this field. Systems have been made which can describe the content of an image, audio or a video. But this is also an example of “Input-output” approach. The tricky part comes when we deal with emotions in such systems. Emotions are not something which we can describe or prove their existence. They can be felt and understood. So how do we make the systems understand emotions? How do we GENERATE emotions in such systems?

Detection of emotions is comparatively easier as this problem also can be broken down into smaller “Input-Output” problems. For example, if we are designing a system capable to detect emotions then it would take input the voice samples, video samples and figure out the emotion/mood of the person based on the voice/gestures/actions etc. The real question is, how do we generate emotions? According to some of the biggest names in the field of AI (Artificial Intelligence) like Stephen Hawking, Andrew Ng and many more, we need a more advanced DEEP LEARNING DATA DRIVEN model for that. If language can be learnt by such systems by observing humans, then it is not impossible to learn emotions by observing humans. This sounds like a very bold and unsupported statement considering today’s situation in this field but we must also keep in mind that what sounded impossible before 20 years is now possible with ease (The famous object detecting algorithm without CNN was able to perform at almost 60% accuracy and after the introduction of CNN it is now almost 99%). New challenges arrive, technologies upgrade and life becomes more and more convenient. This is the trend of the present world. In the end, an open mind always helps in research!

Hypothesis 2

This hypothesis requires a thought experiment. The following experiment is very hard to implement and thus we would just imagine the situation and predict the outcomes. Till now, we have developed systems capable of learning languages from humans. But what would happen if we just create such systems and do nothing!? What would happen if we isolate such systems and give no linguistic input to them? When a child acquires language from its parents, a linguistic input is available to the infant and thus it learns the language spoken by its parents. But consider the case of *Homo-Sapiens* who did not have any language to communicate. What we do know is that the INSTINCT OF SURVIVAL and SELF-REALIZATION were the most important causes of genesis of language in humans. Now, if we try to relate today's intelligent machines with the *Homo-Sapiens,* can there be generation of an all-together new language? The answer to this question is NO! Because today's technology is not sufficient to implement "Self-Realization" or "Instinct of Survival" in machines. But if we look at the massive increase in research in this topic, we can definitely imagine much better versions of intelligent machines in near future. Will genesis of language be possible by such Systems?

Consider a situation where we create say 10 such intelligent systems and isolate them from humans and all forms of life in a virtual environment. All the systems would be equipped with all the necessary items needed for communication (capable to send output and receive input). There are almost infinitely many possibilities what might happen. In earlier hypothesis we have hypothesized the existence of emotions in such systems in future so here we assume the same. If the instinct of fear increases then the systems would simply fight and

destroy each other (As humans would kill each other). If they don't destroy each other there is a possibility that they would try to survive individually without any communication. Even if they try to survive together, there is a possibility that they may not communicate or communicate in sign language. Thus this experiment is a game of possibilities and simulations. Our thought experiment is basically a trial and error game, where we simulate a situation where we force a number of intelligent systems to survive on their own without any linguistic input (language not given). The second hypothesis comments on the outcomes of this experiment.

EVEN THOUGH CHANCES ARE VERY LESS, MAYBE ONCE IN A MILLION TIMES, GENESIS OF A COMPLETELY DIFFERENT LANGUAGE IS POSSIBLE AS AN OUTCOME OF THIS EXPERIMENT IF WE SIMULATE THE EXPERIMENT SUFFICIENT AMOUNT OF TIMES.

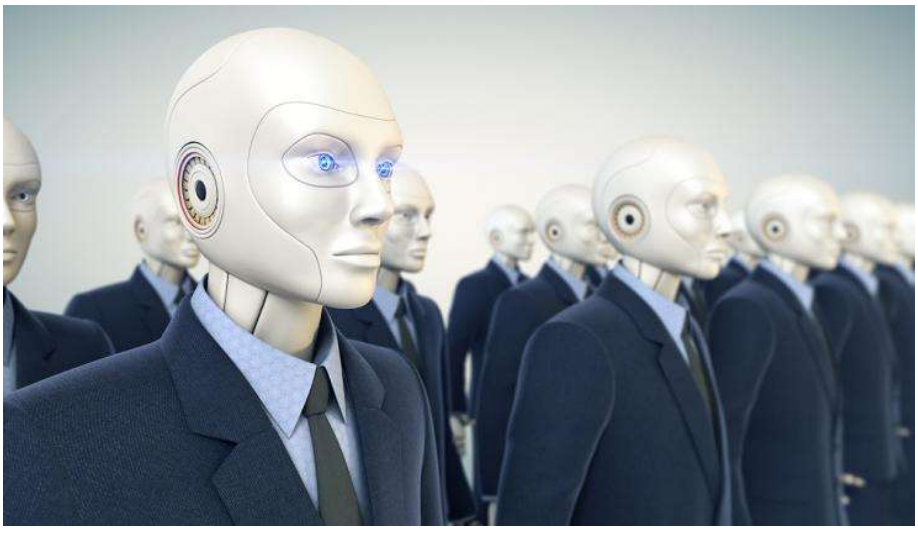
Before making this statement, it is necessary to understand the requirements for generation of a new language. In the experiment we isolate the systems so that no human/living input is given to such systems. This is necessary to generate an authentically new language. In addition to that, we force them to communicate by making such virtual environment where survival is difficult if tried independently. Still there would be cases where the systems would try to survive on

their own and fail. This is acceptable as our systems are Data Driven. In such cases we simulate the experiment once again. The hypothesis states that with a negligible probability there is a chance that the systems would survive together and generate a new language on their own (If they generate a language on their own, it is highly unlikely that they would again generate an already existing language. If they do so then we simply simulate the experiment again and expect better luck!). Thus we see that SELF REALIZATION is very important for genesis of a language. If we look at the genesis of language in humans, we notice that humans didn't know they were capable of speaking /listening /generating a language. It was an instinct in them which drove them to have language. A theory has been proposed stating that Language is a Genetic property of humans and it can't be generated by any other form. But this paper counters this argument as anything can't just appear in Genes of humans on its own. If language is a genetic thing then it must have been introduced in genes at some point in human evolution and "Genetic Changes are always triggered by factors in outside environment like need of survival/ self-realization etc." Thus Genetic change is a consequence of the root cause mentioned above.

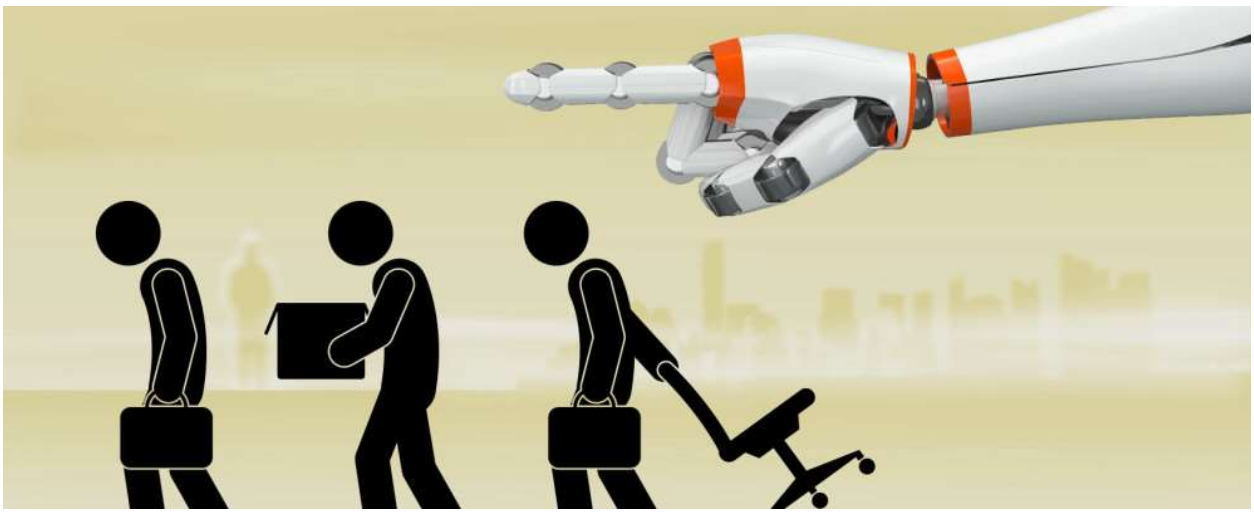
Consequences

Even though these arguments sound futuristic, the referred future is very near, considering the rapid development in the field of AI. If not today, tomorrow these things are going to happen. At first, it all seems very promising and encouraging as machines are basically made to help humans. But the consequences of these developments may harm us.

Consider the case where machines are able to produce literature. If they are capable of producing literary items, it won't take any long to make them produce quality work. Thus there is a possibility that such literary work would be MANUFACTURED and not GENERATED. Imagine a society where such literature is manufactured. Would there be any worth to human-literature in such society? Obviously human-literature would consist of some flaws and would not be able to compete with machine-literature. If too much good literature is available then the Worth of such good literature would decrease. In short, too much technically and emotionally correct literature would spoil the essence of literature. Shakespeare became Shakespeare because he was the only one in a thousand years to achieve that excellence. Worth depends mainly on uniqueness!



If the hypothesis-2 comes out to be true then it would be a major step in making Robots more like humans. Today, humans lack machine qualities like large memory, extended life and fast data processing. On the other hand, machines lack human abilities like diverse thinking, emotions, self-realization etc. But if we try to enable the machines with human qualities, we would be making a kind of new species BETTER than humans in all the aspects. This sounds a bit scary if we understand what exactly are we doing. In process of making our servant better and better, if we make it better than us then it may not serve us anymore. There is a possibility that the human race would be treated as a PEST by the machines and in process of comforting our lives we would eventually destroy our species. Think about the case if in future machines refuse to work for us. Right now also, we are very much dependent upon them for almost everything. If they don't work for us, humans would fail to survive! This statement is not made fictionally but a lot of thinking is present behind it. A serious thought must be given to this possibility otherwise things would spin out of control.



Conclusion

The topic discussed in this paper is indeed a huge one and can be termed “Machine Linguistics” (There might be another name also!). We have briefly discussed about the forms of communication in machines, types of machines applicable for this topic, machine learning, and approaches to tackle linguistic problems in machines, Natural Language Processing and many more. We have also given linguistic predictions in this field and discussed about its consequences.

As we can see, this is a new field and there is a lot scope for research in this field from both the technical and linguistic point of view. There is a necessity to work in this field as this field is very volatile and delicate. Single step taken in this field can have huge consequences on human race and human language. Thus to work in this field, it is required to have a knowledge of both the technology and linguistics. As this topic is largely dependent on AI, it may seem exciting to go deep and expand the scope of machines but as stated in the section “Consequences”, every step must be carefully taken. Excluding the possible fatal outcomes, this field is very interesting and there is much scope to expand. It is hoped that this paper proves to be useful for further research in this topic and taken as a guideline to explore things.

Acknowledgements

There have been many persons and resources due to which this paper came into existence. Firstly, I would like to thank Assistant Professor Suranjana Barua for giving me the opportunity to work on this extremely interesting as well as important topic and guiding me throughout the term as well. Then I would like to thank my brother EESHAN DHEKANE (currently in his 5th year of B.Tech at IIT Kanpur in EEE and CSE department) for introducing all the technical terms used in this paper. I would also like to thanks my parents for supporting and encouraging me. Lastly, I would like to thanks all those who directly or indirectly helped me for this paper (like our librarian who allowed me to take an extra book on Artificial Intelligence). Without this support, it would have been very hard to do such work.

References

1. Basic machine learning algorithms by Andrew Ng (Lecture series at Stanford University).
2. Image processing and object identification by Fei-Fei-Lee and Andrej Karpathy (Lecture series at Stanford University).
3. Introduction to English language and linguistics by Edward, Bernd and George.
4. Language and linguistics- The key concepts by Trask and Stockwell
5. Real world machine learning by Brink, Richards and Fetherolf
6. http://www.andrewng.org/ Andrew Ng's blog
7. https://www.stanford.edu/ Stanford university's website