भाषा प्रौद्योगिकियों में एआई/एमएल की बढ़ती भूमिका

Increasing Role of AI/ML in Language Technologies

महेश कुलकर्णी

Mahesh Kulkarni

Former Sr. Director Corporate R–D, Hod GIST
Former Country manager W3C India
maheshdkulkarni@gmail.com

सारांश

कृत्रिम बुद्धिमत्ता के क्षेत्र में 1966-1997 के दौरान सर्दियों की अवधि देखी गई और इस क्षेत्र ने प्रचार चक्रों के साथ-साथ निराशा और आलोचना का अनुभव किया। कम फंडिंग ने प्राकृतिक भाषा प्रसंस्करण (एनएलपी), और विशेष रूप से मशीन अनुवाद प्रणाली और भाषण प्रौद्योगिकियों में अनुसंधान को प्रभावित किया।

हालाँकि, GPU कंप्यूटिंग के आगमन और बड़े पैमाने पर लेबल किए गए डेटा सेट ने प्राकृतिक भाषा प्रसंस्करण में अनुसंधान को उत्प्रेरित किया।

दुनिया भर में उपयोग की जाने वाली 7000 भाषाओं और बोलियों के साथ, नेटवर्क पर अगले एक अरब उपयोगकर्ताओं को लाने के लिए बहुभाषावाद अत्यंत महत्वपूर्ण पहलू बन गया है।

Abstract:

The field ofArtificial Intelligence w a winter period during 1966–1997 and the sector experienced hype cycles as well as dippointment and criticism. Low funding ffaected research in Natural Language Processing (NLP), and in particular Machine Translation Systems and Speech Technologies.

However, the advent of GPU computing and large–scale labelled data sets catalysed research into Natural Language Processing.

With 7000 languages and dialects used around the world, multilingualismhas become an extremely important aspect of bringing the next billion users to the network.

संक्षेपाक्षर

AI	-	एआई - आर्टिफिशियल इंटेलिजस
CAT	-	कैट - कंप्यूटर असिस्टेड ट्रांसलेशन
CERT–In	-	सीईआरटी - इन - भारतीय कंप्यूटर आपातकालीन प्रतिक्रिया दल
DARPA	-	डारपा - रक्षा उन्नत अनुसंधान परियोजना एजसी
GEM	-	जेम - गवर्नमट ई-मार्केटप्लेस
GPU	-	जीपीयु - ग्राफिकल प्रोसेसिंग यूनिट
GSTN	-	जीएसटीएन - माल और सेवा कर नेटवर्क

INSCRIPT	-	इन्स्क्रिप्ट - भारतीय लिपि
IPR	-	आईपीआर - बौद्धिक संपदा अधिकार
IT	-	आईटी - सूचना प्रौद्योगिकी
MAT	-	एमएटी - मशीन असिस्टेड ट्रांसलेशन
MCA21	-	एमसीए 21 - कॉर्पोरेट मामलों के मंत्रालय
ML	-	एमएल - मशीन लर्निंग
NLG	-	एनएलजी - प्राकृतिक भाषा पीढ़ी
NLP	-	एनएलपी - प्राकृतिक भाषा प्रसंस्करण
NLU	-	एनएलयू - प्राकृतिक भाषा समझ
S2S	-	एसटुएस - स्पीच टू स्पीच
SR	-	एसआर - भाषण पहचान
TDIL	-	टीडीआईएल - भारतीय भाषाओं में प्रौद्योगिकी विकास
TTS	-	टीटीएस - टेक्स्ट टू स्पीच सिस्टम
UPL	-	युपीआय - एकीकृत भुगतान इंटरफ़ेस

भाषा प्रौद्योगिकियों में एआई (AI)/एमएल (ML) की बढ़ती भूमिका

मशीन लर्निंग में प्रगति ने प्राकृतिक भाषा प्रसंस्करण (एनएलपी), आर्टिफिशियल इंटेलिजस के क्षेत्र में अनुसंधान को उत्प्रेरित किया है जो कंप्यूटर को मानव भाषा को समझने की क्षमता देता है। हमने चौटबॉट्स, सर्च इंजन में ऑटोकंप्लीट, वॉयस असिस्टट, लैंग्वेज ट्रांसलेटर, सटीमट एनालिसिस, ग्रामर चेकर्स, डॉक्यूमट क्लासिफिकेशन और फ़िल्टरिंग और अन्य जैसे विभिन्न एनएलपी अनुप्रयोगों के साथ, आई की शक्ति का अनुभव करना शुरू कर दिया है।

डिजिटल क्रांति ने वास्तव में दुनिया को एक विश्व ग्राम बना दिया है। इसमें कृत्रिम बुद्धि यानी आर्टिफिशियल इंटेलिजस का प्रयोग व्यापक होता जा रहा है और यह हमारे दैनिक जीवन का एक अभिन्न अंग बनता जा रहा है।

आप प्रसिद्ध एलन टयूरिंग टेस्ट के बारे में अवगत होंगे जिसमें कहा गया है कि -

''एक कंप्यूटर बुद्धिमान कहलाने के योग्य होगा यदि वह मनुष्य को यह विश्वास करने के लिए धोखा दे सकता है कि वह मानव है।''

कृत्रिम बुद्धि का पूरा सरगम मनुष्यों को 'समझने' और 'प्रतिक्रिया' करने में सक्षम मशीनों पर आधारित है। प्रौद्योगिकी की दुनिया में, मंत्र ''नवप्रवर्तन करो या मरोष्य संगठनों के लिए पहले से कहीं अधिक सत्य है, और आर्टिफिशियल इंटेलिजस (AI) उपयोगकर्ताओं को अधिक वैयक्तिकरण प्रदान करके, प्रक्रियाओं को स्वचालित करके और हमारे काम करने के तरीके को बाधित करके उद्योगों को फिर से परिभाषित कर रहा है। सौभाग्य से, कई कंपनियां।'' तकनीक को अपनाने की महत्ता समझ रही हैं, ऐसा न हो कि वे पीछे रह जाएं।'' एआई को गले लगाओ, या समाप्त जो जाओ, हर संगठन के लिए सच है।

आज, एआई व्यापक है और यह आवाज, छवि पहचान, खोज इंजन, सिफारिश इंजन, स्वचालित ईमेल और पाठ उत्तर, चौटबॉट और बहुत कुछ को शक्ति प्रदान करता है। एआई-फर्स्ट गूगल का नया मंत्र है।

क्षेत्र जहां एआई का सबसे अधिक प्रभाव हो सकता है -

- स्वास्थ्य देखभाल
- कृषि
- शिक्षा
- स्मार्ट सिटी और इंफ्रास्ट्रक्चर
- स्मार्ट गतिशीलता और परिवहन

नीति आयोग ने कृत्रिम बुद्धि (AI) के लिए एक संपूर्ण राष्ट्रीय रणनीति तैयार की है। विभिन्न क्षेत्रों में एआई के व्यापक अनुकूलन के लिए, और विशेष रूप से ई-गवर्नस डोमेन में निम्नलिखित बिंदुओं पर विचार करने की आवश्यकता है–

- **प्रौद्योगिकी की तैयारी और इससे जुड़ी चुनौतियों का समाधान :** एआई के अनुसंधान और अनुप्रयोग में व्यापक-आधारित विशेषज्ञता का अभाव।
- **स्किल मैनपावर जनरेशन :** बड़ी कंपनियों के 54% कर्मचारियों को 2025 तक काम पर बने रहने के लिए कौशल (स्किल) बढ़ाने की जरूरत होगी 2018 फोर्ब्स की रिपोर्ट में दावा किया गया है।
- **नीति / कानूनी ढांचा :** 'स्पष्टीकरण के अधिकार' के लिए अर्थात, व्याख्या करने योग्य AI, GDPR (सामान्य डेटा संरक्षण व्यवस्था), IPR / पेटट व्यवस्था।

- **डेटा उपलब्धता :** (डेटा सुरक्षा, गुमनामी, सत्यापन, सुरक्षित, कानूनी ढांचा / उल्लंघन के लिए कानून)

डेटा नई अर्थव्यवस्था है और नवाचार को उत्प्रेरित कर रहा है और आर्टिफिशियल इंटेलिजस और मशीन लर्निंग वर्ल्ड में व्यवधान लाता है। विभिन्न मिशन मोड परियोजनाओं और यूपीआई, जीएसटीएन, जीईएम, ई-कोर्ट, स्वास्थ्य, एमसीए21 जैसे विभिन्न सार्वजनिक डिजिटल प्लेटफार्मों के माध्यम से विशाल डेटा उत्पन्न हो रहा है। डेटा की विविधता में शामिल हैं - भाषण, पाठ डेटा, वीडियो, छवि, भाषायी संग्रह। 36%AI एप्लिकेशन छवि-आधारित हैं, और कुल मिलाकर, 73%AI एप्लिकेशन किसी न किसी रूप में छवि, वीडियो, ऑडियो या ससर डेटा के साथ काम करते हैं।

सभी रास्ते डेटा की ओर ले जाते हैं– डाटा को उपयोग में लाना, उसे उपयोगी बनाना आसान नहीं है। डेटा उपलब्ध होना चाहिए, इससे ज्ञान निकालने के लिए ड्रिल डाउन किया जाना चाहिए।

- गोपनीयता और सुरक्षा
- डेटा की गुमनामी के बारे में औपचारिक नियम
- बड़े संगठित स्वच्छ डेटा की उपलब्धता, कद्र राज्य सरकार - डेटा एपीआई (डेटा एकत्रीकरण मंच)
- डेटा सुरक्षा कानून, डेटा साझाकरण नीति, सुरक्षित डेटा एक्सेस, गुमनामी

समयरेखा : अब AI क्यों?

1966-1997 को एआई शीतकालीन अवधि कहा जाता है – जिसने एआई अनुसंधान में कम धन और रुचि का अनुभव किया गया। इस क्षेत्र ने प्रचार चक्र और निराशा और आलोचना का अनुभव किया, जिसके बाद धन में कटौती हुई। कई विफलताएँ जैसे 1966 मशीन अनुवाद की विफलता, 1970 : मानव संज्ञानात्मक कौशल समझ का परित्याग, 1971-75 : एआई शीतकालीन की अवधि के लिए जिम्मेदार भाषण समझ के साथ DARPA की निराशा। लेकिन फिर से, इस क्षेत्र में दशकों के बाद नए सिरे से दिलचस्पी दिखाई गई है। कंप्यूटर कुछ संकीर्ण कार्यों में हमसे बेहतर प्रदर्शन कर सकते हैं; हालाँकि, ऐसा महसूस किया जाता है कि मानव-समान A.I. 5 से 10 साल में उभरेगा।

दूसरी ओर, इंटरनेट व्यापक हो गया है और हमारे जीवन का एक अभिन्न अंग बन गया है। हम सभी ने विशेष रूप से महामारी में इंटरनेट की शक्ति को देखा है, जिसने हमें हमेशा की तरह जुड़े रहने के साथ-साथ व्यापार करने में मदद की। जनवरी 2021 तक, वैश्विक सक्रिय इंटरनेट उपयोगकर्ता 4.66 अरब है, जबकि इंटरनेट की 90-95% खपत अकेले सोशल नेटवर्किंग के उपयोग से है।

AI (आर्टिफिशियल इंटेलिजस) और IOT के अभिसरण ने उद्योगों, व्यवसाय और अर्थव्यवस्थाओं के कार्य करने के तरीके को फिर से परिभाषित किया है। स्पीच टू स्पीच टेक्नोलॉजी, फेशियल रिकग्निशन, वर्चुअल असिस्टट, मशीन ट्रांसलेशन सिस्टम, नेचुरल लैंग्वेज प्रोसेसिंग, नेचुरल लैंग्वेज जेनरेशन और बहुत कुछ अब हमारे जीवन का हिस्सा बन रहे हैं, और भाषा की बाधाओं को दूर करने में मदद करते हैं।

2022 में 50 अरब डिवाइस कनेक्ट होने की उम्मीद है, 50 वर्षों के भीतर हमारे पास इंटरनेट ट्रांसीवर को मानव मस्तिष्क में एम्बेड करने की तकनीक होगी और 2069 तक ब्रेन-मशीन इंटरफ़ेस पूरी तरह से विकसित हो जाएगा, जिसमें इंटरनेट इको सिस्टम मानव उन्नति के लिए उत्प्रेरक होगा। आवासीय इंटरनेट की गति 10 गीगाबिट प्रति सेकंड - आज के नेटवर्क की तुलना में 10 गुना तेज होगी।

नेटवर्क पर अगले एक अरब उपयोगकर्ताओं को लाने के लिए बहुभाषावाद अत्यंत महत्वपूर्ण पहलू बन गया है। दुनिया भर में 7,000 भाषाओं और बोलियों का इस्तेमाल किया जाता है। भारत में हमारे पास 22 अनुसूचित भाषाएं हैं, और लिपियों और भाषाओं के बीच हमारे पास एक से कई और कई से कई संबंध हैं। एक उदाहरण के रूप में, देवनागरी लिपि अकेले 9 अनुसूचित भाषाओं को शामिल करती है, जैसे कि हिंदी, मैथिली, मराठी, कोंकणी, बोरो, नेपाली, संताली, संस्कृत, सिंधी, जबकि सिंधी देवनागरी के साथ-साथ फारसी-अरबी लिपि में भी लिखी जाती है।

अगले अरब इंटरनेट उपयोगकर्ता संभवतः गैर-अंग्रेज़ी भाषी देशों से आएंगे, इन उपयोगकर्ताओं के लिए पहुँच

प्रदान करने के लिए अंतर्राष्ट्रीयकृत या बहुभाषी सामग्री का समर्थन करने से अधिक की आवश्यकता होगी। स्थानीयकृत डोमेन नाम और ईमेल पते आवश्यक हैं।

उपभोग और साथ ही बहुभाषी सामग्री का निर्माण भी बढ़ रहा है, जो मानव प्रेरक प्रणालियों में प्रगति के लिए एक वरदान है। मशीन लर्निंग में प्रगति ने प्राकृतिक भाषा प्रसंस्करण (एनएलपी), आर्टिफिशियल इंटेलिजस के क्षेत्र में उल्लेखनीय प्रगति की है जो कंप्यूटर को मानव भाषा को समझने की क्षमता देता है।

क्या बदला है? अब क्यों

परिवर्तन के महत्वपूर्ण घटक बड़े पैमाने पर लेबल किए गए डेटा सेट और GPU कंप्यूटिंग, बेहतर आर्किटेक्चर / एल्गोरिदम, सॉफ़्टवेयर प्लेटफ़ॉर्म—Tensor Flow, Theano, Chainer और MXnet जैसे फ्रेमवर्क की उपलब्धता है। आर्टिफिशियल इंटेलिजस का उपयोग करने वाले अधिकांश समाधान प्राकृतिक भाषा समझ (एनएलयू) और इमेज प्रोसेसिंग और कंप्यूटर विजन (IPCV-आईपीसीवी) पर लागू होते हैं।

भाषा (प्राकृतिक भाषा समझ और निर्माण)

68% लोग गाँवों में रहते हैं, 22 अनुसूचित भाषाएँ, विभिन्न बोलियां, केवल 7-8 प्रतिशत अंग्रेजी समझ सकते हैं, 36ण्5% लोग अपनी भाषा पढ़/लिख नहीं सकते हैं, लेकिन केवल बोल सकते हैं। हम स्पीच टू स्पीच टेक्नोलॉजीज (S2S) जैसे ह्यूमन इंस्पायरिंग सिस्टम की आवश्यकता है।

मशीन लर्निंग ने प्राकृतिक भाषा प्रसंस्करण (एनएल. पी), आर्टिफिशियल इंटेलिजस के क्षेत्र में अनुसंधान को उत्प्रेरित किया है जो कंप्यूटर को मानवी भाषा को समझने की क्षमता देता है। प्राकृतिक भाषा लोगों द्वारा बोली जाने वाली और लिखित और संचार के लिए उपयोग की जाने वाली भाषा को संदर्भित करती है, जबकि एनएलपी एल्गोरिदम का उपयोग करके शब्दों और वाक्यों से जानकारी निकालती है। इसके अलावा नेचुरल लैंग्वेज जनरेशन (एनएलजी) मानव जैसे वाक्यांशों को तैयार करने की क्षमता प्रदान करता है, और प्राकृतिक भाषा समझ (एनएलयू), वाक्यांशों की समझ बनाने की क्षमता प्रदान करता है।

एनएलयू उपकरण प्राकृतिक भाषा में पाठ या आवाज को संसाधित करते हैं और प्रतिक्रियाओं को सारांशित, संपादित बनाकर उपयुक्त करते हैं।

एनएलयू अनुसंधान के कुछ क्षेत्र निम्नलिखित हैं :

- स्पीच टू स्पीच टेक्नोलॉजी (टीटीएस, एमएटी, एसआर)।
- प्रश्न-उत्तर प्रणाली, चौटबॉट।
- सूचना निष्कर्षण और सूचना पुनर्प्राप्ति प्रणाली
- मुद्रित और हस्तलिखित पहचान।
- प्राकृतिक भाषा समझ - खोज, प्रश्न-उत्तर, चौटबॉट, प्रवचन-विश्लेषण, भावना-विश्लेषण, सोशल मीडिया-विश्लेषण, आदि।

डिजिटल माध्यम में भाषा अनुकूलन

भाषा(ओं) के साथ किसी भी उपकरण/प्रणाली को सक्षम करने के लिए मूल रूप से चार घटकों की आवश्यकता होती है।

- डेटा कैसे इनपुट कर।
- डेटा कैसे स्टोर कर
- डेटा कैसे प्रदर्शित कर
- और बाहरी दुनिया के साथ संवाद कैसे कर।

प्रमुख रूप से, इन्हें भारतीय मानकीकृत कीबोर्ड लेआउट (उन्नत INSCRIPT), भंडारण के लिए यूनिकोड और प्रदर्शन के लिए ओपन फॉन्ट प्रारूप का उपयोग करके हल किया जा रहा है, फिर भी गहरे स्तर के समर्थन में चुनौतियां हैं, जैसे कि डेटाबेस में भाषा समर्थन। हालांकि सभी डेटाबेस यूनिकोड एन्कोडिंग का समर्थन करते हैं, लेकिन भाषा समर्थन से संबंधित गहरे मुद्दे आज भी मौजूद हैं। जैसे कि एकाधिक भाषाओं में डेटाबेस क्वेरी, होमोफ़ोनिक विविधताएं, वर्तनी भिन्नताएं, वैकल्पिक वर्तनी, सॉर्ट ऑर्डर और कई अन्य भाषायी विशेषताएं। कुछ डेटाबेस विक्रेताओं ने भाषायी (भाषा विशिष्ट) कुछ ही विविधताओं (उदाहरण गीता और गिता - Geeta, Gita) के समर्थन के लिए सक्रिय उपाय किए हैं, लेकिन सभी को नहीं। इसलिए, संबंधित

डेटाबेस विक्रेताओं द्वारा ऐसी भाषायी विशेषताएं की चुनौतियों पर ध्यान देने की आवश्यकता है।

एक और मुद्दा, लैटिन फॉन्ट की प्रचुरता की तुलना में कम अच्छी गुणवत्ता वाले टाइपोग्राफिक विकल्प हैं, विशेष रूप से भारतीय भाषाओं की लिपियों के लिए। हालाँकि, ई-गवर्नस अनुप्रयोगों के लिए सरकार। टीडीआईएल, एमईआईटीवाई, सरकार के तहत विकसित ष्सकल भारतष् फ़ॉन्ट के उपयोग की सिफारिश करता है। इसके साथ ही, स्टार्ट-अप, फॉन्ट के विकास के लिए खुद को संलग्न कर सकते हैं और उसी के आसपास व्यवसाय उत्पन्न कर सकते हैं। फॉन्ट बिजनेस में बहुत बड़ा स्कोप है।

भारतीय भाषाओं के संबंध में अनुवाद और अन्य भाषायी उपकरणों के मुद्दों की एक बड़ी संख्या है। हालांकि, भाषायी उपकरण जैसे वर्तनी जांचकर्ता, व्याकरण परीक्षक और मशीनी अनुवाद अत्यधिक शोध उन्मुख हैं और किसी विशेष डोमेन के संदर्भ में ठीक से व्यवहार करगे। एक उदाहरण के रूप में, सामान्य शब्दों के लिए वर्तनी जांचकर्ता कानूनी या चिकित्सा डोमेन के लिए काम नहीं कर सकता है। ऐसा ही मामला मशीनी ट्रांसलेशन सिस्टम का है। तो, सभी के लिए उपयुक्त एक समाधान नहीं होगा। इसलिए, यहां भी स्टार्ट-अप के लिए इन तकनीकों में खुद को शामिल करने और उसी के व्यवसाय पर पनपने का एक जबर्दस्त अवसर है। दूसरे, जटिलता के कारण और यहां तक कि मानव द्वारा अनुवादित दो वाक्य भी समान नहीं होंगे, Man in loop की आवश्यकता है। इसलिए, स्टार्ट-अप पोस्ट एडिटिंग के लिए Man in loop के साथ अनुवाद का काम करने के लिए खुद को संलग्न कर सकते हैं। संपूर्ण अनुवाद उद्योग CAT टूल पर फल-फूल रहा है।

सरकारी विभागों को भारतीय स्टार्ट-अप उत्पादों के पहले उपयोगकर्ता बनने के लिए प्रोत्साहित किया जाना चाहिए ताकि एक बड़ा इंडिक डाटासेट उपलब्ध/उत्पन्न हो, जो बाद में एक बृहद डेटाबेस की ओर ले जाए।

प्राकृतिक भाषा प्रसंस्करण क्षेत्र में विश्व का नेतृत्व करने की क्षमता भारत में है, इसलिए 'एनएलपी फर्स्ट' हमारा दृष्टिकोण होना चाहिए। हम निम्नलिखित बातों पर ध्यान देने की आवश्यकता है :

- एआई/एमएल अनुसंधान में उपयोग के लिए डेटा सहित भाषायी संसाधनों के सृजन के लिए राज्य सरकार की भूमिका में वृद्धि।
- भारत सरकार की सभी सेवाओं, सूचनाओं (राज्य/कद्र) को स्थानीय भाषाओं में बनाना।
- अनिवार्य रूप से भारत सरकार की सभी वेबसाइट सभी 22 भारतीय भाषाओं में होनी चाहिए।
- स्थानीय भाषा समर्थन के संबंध में एक नीति होनी चाहिए और डिफ़ॉल्ट रूप से भारत सरकार की सभी / कोई भी वेबसाइट स्थानीय भाषा में होनी चाहिए।
- यदि 22 भाषाओं में नहीं तो कम से कम त्रिभाषा सूत्र नीति को अपनाया जा सकता है, अर्थात अंग्रेजी, हिंदी और स्थानीय भाषा।
- अनुवाद/स्थानीयकरण की गतिविधि को चलाने के लिए कुछ प्रतिशत विभागों के आईटी बजट को स्थानीयकरण के लिए निर्धारित किया जा सकता है।
- सीईआरटी - इन द्वारा पैनल में शामिल सूचना सुरक्षा ऑडिटिंग संगठनों की तर्ज पर पैनल में शामिल अनुवाद एजसियों को रखा जा सकता है।

In 1972, Mr. Fritjof Capra, an American scientist had written in detail about the relation between the dance of sub-atomic particles and the cosmic dance of Shiva that "Modern physics has shown that the rhythm of creation and destruction is not only manifest in the turn of the seasons and in the birth and death of all living creatures, but is also the very essence of inorganic matter", and that "For the modern physicists then, Shiva's dance is the dance of subatomic particles".