BOOMI INTERNATIONAL RESEARCH JOURNAL OF TAMIL
E - ISSN: 3139 -1893
An international open access, peer- reviewed, refereed journal

இயற்கை மொழி புரிதலில் (NLU) இயந்திரக் கற்றல் (ML)

Machine Learning (ML) in Natural Language Understanding (NLU)

முனைவர் ஆர். சண்முகம்
செயல்பாட்டு மேலாளர்– ஜியோ பிளாட்ஃபார்ம்ஸ் லிமிடெட், பெங்களூரு – 560025, கர்நாடகா, இந்தியா
Dr R. Shanmugam
Operations Manager – Jio Platforms Limited, Bengaluru – 560025, Karnataka, India
Email: shanfaace@gmail.com

Introduction

When you tell a voice assistant, "Book a taxi to the airport at seven in the morning," several things must happen before it can actually act. First thing is It must identify the user's intent (booking a taxi), extract the destination (the airport), note down the time (7:00), and understand whether "seven" means seven in the morning or seven in the evening. This is exactly the task of Natural Language Understanding (NLU). it can be thought of as the branch of natural language processing that extracts structured meaning from unstructured data or speech.

In the past, this was handled using the grammar rules and root dictionaries. That approach still remains useful for closed domains and morphologically poor languages. But modern NLU systems increasingly rely on Machine Learning. This article outlines how Machine Learning contributes to NLU, which techniques matter, and the challenges practitioners face along the way.

Keywords : Natural Language Understanding, Machine Learning, Natural Language Processing, BERT

ஆய்வுச் சுருக்கம்

"காலை ஏழு மணிக்கு விமான நிலையத்திற்கு ஒரு டாக்ஸி புக் பண்ணு" என்று நீங்கள் ஒரு குரல் உதவியாளரிடம் சொல்லும்போது, அது செயல்படும் முன்பு பல விஷயங்கள் நடக்க வேண்டும். ஒரு பயனரின் நோக்கத்தை (டாக்ஸி புக் செய்தல்) அடையாளம் காண வேண்டும், சேருமிடத்தைத் (விமான நிலையம்) பிரித்தெடுக்க வேண்டும், நேரத்தைத் (7:00) துல்லியப்படுத்த வேண்டும், மேலும் "ஏழு" என்பது காலை ஏழு மணியா அல்லது இரவு ஏழு மணியா என்பதைப் புரிந்து கொள்ள வேண்டும். இதுவே இயற்கை மொழி புரிதலின் (Natural Language Understanding) பணியாகும் — கட்டமைப்பற்ற தரவுகளிலிருந்து அல்லது பேச்சிலிருந்து கட்டமைப்பான பொருளைப் பிரித்தெடுக்கும் இயற்கை மொழி செயலாக்கத்தின் ஒரு பிரிவாக இதைக் கருதலாம்.

முன்பெல்லாம் இது இலக்கண விதிகள் மற்றும் அகராதிகளைப் பயன்படுத்திச் செய்யப்பட்டது. குறுகிய துறைகள் மற்றும் உருபனியல் வளம் குறைந்த மொழிகளுக்கு அந்த அணுகுமுறை இன்னும் பயனுள்ளதாகவே உள்ளது. ஆனால் நவீன இ.மொ.பு (NLU) அமைப்புகள் பெருகிய முறையில் இயந்திரக் கற்றலை (Machine Learning) நம்பியிருக்கின்றன. இந்தக் கட்டுரை, இ.மொ.பு (NLU)-க்கு இயந்திரக் கற்றல் எவ்விதத்தில் பயனளிக்கிறது, மேலும் இதில் எவ்வித நுட்பங்கள் முக்கியமானவை, இதனோடு இதில் நிபுணர்கள் எதிர்கொள்ளும் சவால்களையும் கோடிட்டுக் காட்டுகிறது.

கலைச்சொற்கள் :இயற்கை மொழி புரிதல், இயந்திரக் கற்றல், இயற்கை மொழியாய்வு, பெர்ட் (BERT)

முன்னுரை

'யாமறிந்த மொழிகளிலே தமிழ்மொழிபோல் இனிதாவது எங்கும் காணோம்' என்ற பாரதியின் தேசியகீத வரிகளுக்கேற்ப தமிழும் தன்னை அறிவியல் மாற்றங்களுக்கு ஏற்றவாறு தகைவமைக்கும் பண்பை இயல்பாகவே பெற்றிருக்கிறது. ஆம் ஒரு மொழியின் இலக்கணம் தேர்ந்த கணிதப்பண்புகளை உள்ளடக்கி இருக்கும் போது அம்மொழியைக் கணினி புரிந்துகொள்வது என்பது சற்று எளிமையாகிறது. வளர்ந்து வரும் மொழித்தொழில்நுட்ப அறிவியலில் கணினி தானாகவே மொழியைப் புரிந்து கொள்ளும் 'இயற்கை மொழி புரிதல்' (இ.மொ.பு) என்ற துறையையும் அதனை இயந்திரக்கற்றல் (ML) வழி எவ்வாறு நிகழ்த்த முடியும் என்பதைப் பற்றியும் சில அடிப்படைகளை அலசுவதே இக்கட்டுரையின் சாரம்சம்.

இ.மொ.பு உண்மையில் என்ன செய்கிறது?

எந்த மொழி அடிப்படையிலான பயன்பாட்டிலும் இ.மொ.பு (NLU) என்பது வாக்கியப் பொருளை உணருவதாகும். பேச்சு அறிதல், ஒலி அலைகளை உரையாக மாற்றுகிறது; உரை உருவாக்கம் புதிய வாக்கியங்களை உருவாக்குகிறது. இ.மொ.பு (NLU) இவற்றுக்கு நடுவே அமர்ந்து பயனர் சொன்னதை இயந்திரத்தால் செயல்பட முடிந்த ஒன்றாக மாற்றுகிறது. நடைமுறையில் இது எவ்வாறு சாத்தியமாகிறது?

- நோக்கம் வகைப்படுத்தல் — பயனர் என்ன செய்ய விரும்புகிறார் என்பதை அடையாளம் காணுதல்.
- உள்ளுறுப்பு பிரித்தெடுத்தல் — பேசப்பட்ட குறிப்பிட்ட விவரங்களை (நேரம், இடம், பெயர்) உள்வாங்கல்.
- இட நிரப்புதல் — அந்த உள்ளுறுப்புகளை அவற்றின் பணிகளுடன் இணைத்தல் ("சென்னை" புறப்படும் இடமா, சேரும் இடமா). இதை ஆங்கிலத்தில் Annotation என்பர்.
- உணர்வு அறிதல் — பயனர் மகிழ்ச்சியாகவா, கோபமாகவா, அல்லது தகவல் கேட்கிறாரா என்பதைக் கண்டறிதல். இதை ஆங்கிலத்தில் Sentiment Analysis என்பர்.

இயந்திரக் கற்றல் ஏன் முக்கியம்

விதிகள் அடிப்படையிலான அமைப்புகள் மொழி கணிக்கக்கூடிய வகையில் இருந்தால் நன்றாகச் செயல்படுகின்றன. ஆனால் மொழி அப்படி இருப்பதில்லை. ஒரு பயனர் "எனக்கு ஒரு விமான டிக்கெட் புக் பண்ணு," "டெல்லிக்கு போக வேண்டும்," "நாளை டெல்லி விமானம்," "டெல்லி ஃபிளைட் ப்ளீஸ்" என்று பலவிதமாகச் சொல்லலாம். ஒவ்வொரு சாத்தியமான வடிவத்திற்கும் விதி எழுதுவது என்பது அரிது. இயந்திரக் கற்றல் இதற்கென வேறு அணுகுமுறையைக் கையாள்கிறது. விதிகளைக் குறிப்பிடுவதற்குப் பதிலாக, ஆயிரக்கணக்கான உதாரணங்களை முன்னரே படித்துப் புரிந்து கொண்டு அவற்றிலிருந்து வாக்கிய வடிவங்களைக் கற்றுக்கொள்கிறது; இதனை ஆங்கிலத்தில் Training data என்பர். பயிற்சித் தரவு போதுமான தரவுகளை உள்ளடக்கினால், அது முன்பு பார்த்திராத சொற்றொடர்களையும் கையாளும் திறன் பெறும். இதனை ஆங்கிலத்தில் Unseen data handling என்பர். இந்தப் பொதுமைப்படுத்தல்தான் இதன் அடிப்படை பலமாக உள்ளது.

இயந்திரக் கற்றலின் நுட்பங்கள்

இ.மொ.பு (NLU)-வில் இயந்திரக் கற்றல் வெவ்வேறு பணிகளுக்குப் பயன்படுத்தப்படுகிறது. நோக்கம் மற்றும் வகைப்படுத்தலென அந்தப் பட்டியல் நீளும், ஆரம்பகால இயந்திரக் கற்றல் முறைகள் நேய்வ் பேயிஸ் (Naive Bayes), எஸ்விஎம் (SVM), லாஜிஸ்டிக் ரிக்ரெஷன் (Logistic Regression) போன்ற எளிய கோட்பாடுகளைக் (Formalisms) கொண்டு பயன்படுத்தின (Sebastiani, 2002). நவீன அமைப்புகள் பெர்ட் (BERT) போன்ற டிரான்ஸ்ஃபார்மர் அடிப்படையிலான மாதிரிகளைப் பயன்படுத்துகின்றன (Devlin et al., 2019). இவை

சூழலைக் கவனித்து அதை நினைவில் வைத்திருக்கும் திறன் கொண்டவை. "புத்தகம் படி" என்பதிலும் "விமானம் புக் பண்ணு" என்பதிலும் உள்ள "புக்" என்ற வார்த்தை வெவ்வேறு பொருள் கொள்கிறது என்பதை இதன் மூலம் கணினி புரிந்துகொள்ளும்.

உள்ளுறுப்பு பிரித்தெடுத்தல் என்பது ஒரு இலக்கணக் குறிப்பறிதல் பணி. இதில் ஒவ்வொரு சொல்லுக்கும் ஒரு இலக்கணக்குறிப்பு (லேபிள்) ஒதுக்கப்படும். முன்பெல்லாம் இலக்கண விதிகள் இதற்காகப் பயன்படுத்தப்பட்டன. ஆழ் கற்றல் அணுகுமுறைகள் (Deep Learning) பிஐஎல்எஸ்டிஎம் (BiLSTM) போன்றவற்றை இதற்காகப் பயன்படுத்தின (Graves & Schmidhuber, 2005); இன்று டிரான்ஸ்ஃபார்மர்கள் முதன்மை வகிக்கின்றன.

சொல் உட்பொதிப்புகள் (word embeddings) என்பது சொற்களை எண் திசைவெளிகளாக மாற்றும் நுட்பம். வேர்ட்2வெக்டர், ஜிஎல்ஓவிஇ (Word2Vec, GloVe) போன்றவை நிலையான உட்பொதிப்புகளை உருவாக்கிக்கொள்ளும் திறனை மாதிரிக்கு வழங்கும் (Mikolov et al., 2013; Pennington et al., 2014). இஎல்எம்ஓ, பெர்ட் (ELMo, BERT) போன்ற சூழல்சார் மாதிரிகள் ஒவ்வொரு சூழலுக்கும் வெவ்வேறு திசைவெளிகளை உருவாக்குகின்றன (Peters et al., 2018; Devlin et al., 2019) — இது இயற்கைமொழி புரிதலை நம் தேவைக்கேற்ப மாற்றிக்கொள்ளப் பயனளிக்கும்.

கடந்த சில ஆண்டுகளில் இ.மொ.பு (NLU)-வில் மிகப்பெரிய மாற்றம் என்பது பெரிய முன்-பயிற்சி பெற்ற மொழி மாதிரிகளின் எழுச்சி (Pre-trained Large Language Models). நம்முடைய ஒவ்வொரு தேவைக்கும் புதிதாக ஒரு மாதிரியைப் பயிற்சி செய்வதற்குப் பதிலாக, பரந்த உரை மூலத்தொகுப்புகளிலிருந்து பொதுவான மொழி வடிவங்களைக் கற்றுக்கொண்ட ஒரு மாதிரியை எடுத்து, உங்கள் குறிப்பிட்ட பணிக்குச் சிறிய லேபிள் தரவுத் தொகுப்புடன் அதை தகவமைத்துப் பயன்படுத்தலாம். இது டிரான்ஸ்ஃபர் லெர்னிங் (transfer learning) என்று அழைக்கப்படுகிறது (Pan & Yang, 2010). இதன் மூலம் ஒரே மொழி மாதிரியை மற்ற மொழித் தரவுகளின்றியும் கூட மற்ற மொழிகளுக்குப் பயன்படுத்தலாம்.

ஒரு எளிய உதாரணம்

இயந்திரக் கற்றல் இ.மொ.பு (NLU)-வில் எப்படிச் செயல்படுகிறது என்பதைக் காண்பிக்க ஒரு எளிய உதாரணத்தைக் காண்போம். "நாளை காலை பத்து மணிக்கு டாக்டர் அப்பாய்ன்ட்மென்ட் பதிவு செய்யுங்கள்" என்று ஒரு பயனர் சொல்கிறார். ஒரு பயனரின் நோக்கம் (Intention) என்கிற வகையில் இதை முன்பதிவு_செய் (Book_Appointment) என்ற நோக்கமாக (Intent) இயந்திரக்கற்றல் அடையாளம் காண்கிறது. ஒரு உள்ளுறுப்பு பிரித்தெடுத்தல் மாதிரி (Entity parsing Model) மூன்று உள்ளுறுப்புகளை வெளியேற்றும்; எடுத்துக் காட்டாக நேரம் (நாளை காலை பத்து மணி), சேவை (டாக்டர் அப்பாய்ன்ட்மென்ட்), மற்றும் செயல் (பதிவு). இந்த இரு பணிகளுக்கும் மாதிரிகள் ஆயிரக்கணக்கான உதாரணங்களிலிருந்து கற்றுக்கொண்ட தனது அறிவைப் பயன்படுத்தும். உதாரணமாக ஒரு புதிய வாக்கியம் — "நாளைக்குக் காலையில பத்து மணிக்கு டாக்டர் புக் பண்ணுங்க" — சரியாக அதே கட்டமைப்பான பொருளைத் தருகிறது, இது பயிற்சி தரவில் அப்படியே இல்லாத போதிலும், பொதுமைப்படுத்தலின் வாயிலாகத் தானாகவே மாதிரியால் இதைக் கையாள முடியும்.

இயந்திரக் கற்றலின் சவால்கள்

இயந்திரக் கற்றல் இ.மொ.பு (NLU)-ஐ செயல்படும் விதத்தை மட்டும் மாற்றியமைத்தது, ஆனால் அனைத்துச் சிக்கல்களையும் இதனால் தீர்க்க இயலாது, முக்கியமாக:

- வளம் குறைந்த மொழிகளுக்கான தரவுப் பற்றாக்குறை. பெரும்பாலான முடிவுகள் ஆங்கிலத்தில் பயிற்சி பெற்ற மாதிரிகளிலிருந்து கிடைக்கப் பெறுகின்றன.

மேலும் இந்திய மொழிகளுக்குத் தேர்ந்த தரவு வளங்கள் மிகவும் குறைவு என்பது இதற்கு ஒரு காரணம்.

- மொழிக் கலப்பு - உண்மையான உரையாடல்களில் பயனர்கள் ஒரே வாக்கியத்திற்குள் மொழிகளை மாற்றிப் பயன்படுத்துகின்றனர். "Book பண்ணுங்க a flight from Chennai to Delhi" போன்றவகை உதாரணங்கள் வெகு இயல்பாக இங்கு பயன்படுத்தப்படுகின்றன. குறிப்பாக ஆங்கிலக் கலப்பு அதிகளவில் பயன்பாட்டில் உள்ளது. பெரும்பாலான நிலையான மாதிரிகள் இதைச் சரியாகக் கையாள்வதில்லை. இதனை ஆங்கிலத்தில் code mixing என்பர்.
- தெளிவின்மை மற்றும் சூழல்- "ஐந்து மணிக்கு அலார்ம் வை" என்பதை வைத்து, காலை ஐந்து மணிக்கா அல்லது மாலை ஐந்து மணிக்கா என்பதை முழுமையாகப் புரிந்து கொள்ள முடியாது.

வாடிக்கையாளர் சேவை தரவில் பயிற்சி பெற்ற மாதிரி மருத்துவ அல்லது சட்ட வினவல்களில் தோல்வியடையக்கூடும். எனவே ஒவ்வொரு துறைக்கும் தனி மாதிரி தயாரிக்கப்படல் அவசியம்.

விதிகள் மற்றும் இயந்திரக் கற்றல் — ஒரு கலப்பு அணுகுமுறை

இது போன்ற பயன்பாடுகளுக்கு, மிகவும் வலிமையான அணுகுமுறை என்பது விதிகளும் அல்ல, இயந்திரக் கற்றலும் அல்ல, இவை இரண்டின் கலவையே ஆகும். நன்கு புரிந்துகொள்ளப்பட்ட துறையின் பகுதிகளை விதிகள் துல்லியமாகக் கையாள்கின்றன; மாறுபாடு மற்றும் எளிதில் கணிக்க முடியாத சூழல்களுக்கு இயந்திரக் கற்றல் உதவுகிறது. குறிப்பாக இந்திய மொழிகள் போன்ற — பெரிய இலக்கணக்குறிப்புத் தரவுத்தொகுப்புகள் இல்லாத — சூழல்களில் இந்தக் கலப்புக் கட்டமைப்பு மிகவும் பயனுள்ளது. mBERT (Pires et al., 2019), XLM-R (Conneau et al., 2020), MuRIL (Khanuja et al., 2021), IndicBERT (Kakwani et al., 2020) போன்ற பன்மொழி முன்-பயிற்சி பெற்ற மாதிரிகள் இதற்கான செயல்திறனின் தோல்வி இடைவெளியை வெகுவாகக் குறைத்துள்ளன, ஆனால் இவற்றை மேம்படுத்த இன்னும் கவனமான தரவுத் தேர்வும், சரியான உள்ளுறுப்பு வடிவமைப்பும், தேவைப்படும் இடங்களில் இலக்கண விதிகளும் தேவை. ஆங்கிலத்தில் இதை Hybrid approach என்பர்.

முடிவுரை

இயந்திரக் கற்றல் இ.மொ.பு (NLU) செய்யக்கூடியதை அடிப்படையிலிருந்து மாற்றியுள்ளது. முன்பு கையால் எழுதப்பட்ட இலக்கண விதிகளால் சமாளிக்கப்பட்ட பணிகளை இப்போது ஒரு சில ஆயிரம் உதாரணங்கள் மற்றும் ஒரு முன்-பயிற்சி பெற்ற மாதிரி மூலம் சமாளிக்க முடியும். ஆனால் இயந்திரக் கற்றல் மொழியியலாளரைச் சாராமல் மேம்படுத்தப்படுவதில்லை; நவீன சிறந்த மாதிரிகள் நெகிழ்வுத்தன்மைக்காக இயந்திரக் கற்றலையும், துல்லியத்திற்காக மொழியியல் அறிவையும் பயன்படுத்துகின்றன. ஒவ்வொன்றையும் எப்போது பயன்படுத்த வேண்டும் என்பதைப் புரிந்துகொள்வதுதான் நம்பகமான இயற்கை மொழி புரிதலை உருவாக்கும் திறமையாகும்.

References:

1. Jurafsky, D., & Martin, J. H. Speech and Language Processing (3rd Edition Draft). Pearson.
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL.
3. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. EMNLP.

4. Sebastiani, F. (2002). Machine learning in automated text categorisation. ACM Computing Surveys, 34(1), 1–47.
5. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.
6. Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM networks. IJCNN.
7. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. NAACL.
8. Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345–1359.
9. Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is Multilingual BERT? ACL.
10. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. ACL.
11. Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., Margam, D. K., Aggarwal, P., Nagipogu, R. T., Dave, S., Gupta, S., Gali, S. C. B., Subramanian, V., & Talukdar, P. (2021). MuRIL: Multilingual Representations for Indian Languages. arXiv:2103.10730.
12. Kakwani, D., Kunchukuttan, A., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020). IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. Findings of EMNLP.