தமிழ் செயற்கை நுண்ணறிவு ஆய்விதழ்
மதிப்பாய்வு செய்யப்படும் இதழ்
5
Tamil Journal of Artificial Intelligence
Peer-Reviewed Journal
தொகுதி - 1, இதழ் - 1, ஜனவரி 2026 மின்-ISSN:

தமிழ் மொழி செயலாக்கத்திற்கு ஊக்கமளிக்கும் AI கருவிகள்

முனைவர் கி. நாகேந்திரன், உதவிப்பேராசிரியர், தமிழ்த்துறை, ஸ்ரீ எஸ்.இராமசாமி நாயுடு ஞாபகார்த்தக் கல்லூரி, சாத்தூர், விருதுநகர் மாவட்டம், தமிழ்நாடு, இந்தியா.
ORCiD: https://orcid.org/0000-0003-2865-7832

அறிமுகம்

மொழி என்பது ஒரு நிலத்தின் பாரம்பரியத்தையும், வாழ்வின் அடையாளத்தையும், பேச்சாற்றல் மற்றும் எழுத்தாற்றலின் தொடக்கத்தையும், ஒரு சமூகத்தின் சுவாசத்தையும் தன்னகத்தே தாங்கி நிற்கும் உன்னதப் படைப்பாகும். தொல்காப்பியம் தோன்றிய காலம் முதலே, தமிழ் மொழியானது தன் இலக்கண நெறிகளோடு இயல்பான வாழ்வுமுறையையும் இணைத்து செழித்து வளர்ந்து வந்துள்ளது எனக் காணலாம். மனிதக் குரலுக்கு இணையான குரலில் கணினிகள் பேசத் தொடங்கியுள்ள இன்றைய நவீன உலகிலும், தமிழ் மொழி தொழில்நுட்பத்தோடு இணைந்து பயணிக்கத் தொடங்கியுள்ளது என்பதை வியப்புடன் பார்க்க நேரிடுகிறது. இப்பயணத்தில், செயற்கை நுண்ணறிவு (Artificial Intelligence) எனப் அழைக்கப்படும் நவீனத் தொழில்நுட்பம் தமிழுக்குப் பல புதுமையான பரிமாணங்களை வழங்கி வருகிறது எண்டது மிகையில்லை. குறிப்பாக, 21-ஆம் நூற்றாண்டில் செயற்கை நுண்ணறிவுத் தொழில்நுட்பம் உலகளாவிய அளவில் மாபெரும் மாற்றங்களை ஏற்படுத்தி வரும் நிலையில், மனித மொழியைக் கணினி புரிந்துகொள்ளச் செய்யும் 'இயற்கை மொழிச் செயலாக்கம்' (Natural Language Processing - NLP) முக்கியப் பங்கு வகிக்கிறது என்பது உண்மையே. இத்துறையில் தமிழ் மொழியின் தற்போதைய நிலை, அதனை ஆவணக்காப்பு செய்யும் செயற்கை நுண்ணறிவுக் கருவிகள் குறித்து இக்கட்டுரையில் விரிவாகக் காண்போம்.

இயற்கை மொழிச் செயலாக்கமும் தமிழ் மொழியும்

இயற்கை மொழிச் செயலாக்கம் (NLP) என்பது மனித மொழியைக் கணினி புரிந்துகொள்ளவும், பகுப்பாய்வு செய்யவும், தகுந்த பதிலளிக்கவும் உதவும் ஒரு அதிநவீனத் தொழில்நுட்பம் சார்ந்த துறையாகும். இது இயந்திர மொழிபெயர்ப்பு (Machine Translation), உரை அடிப்படையிலான தேடல் (Text-based Search), உரையாடல் அமைப்புகள் (Conversational Systems) மற்றும் ஒலி அடிப்படையிலான மொழிப் புரிதல் (Speech-based Language Understanding) ஆகியவற்றை உள்ளடக்கியதாகும். மிகவும் தொன்மை வாய்ந்த தமிழ் மொழி, அதன் செறிவான இலக்கண அமைப்புகள், வேர்ச்சொற்களின் ஆழம், பன்முகப் பயன்பாடுகள் மற்றும் பண்டைய இலக்கிய வடிவங்கள் ஆகியவற்றின் காரணமாக இயற்கை மொழிச் செயலாக்கத்தில் தனித்துவமான சிறப்புகளையும் அதேவேளையில் சில சவால்களையும் கொண்டுள்ளது. தமிழ் மொழியின் பல்வேறு வட்டார வழக்குகள், சங்க இலக்கிய மரபுகள் மற்றும் வேர்ச்சொல் அடர்த்தி ஆகியவை கணினிப் பரிமாணத்தில் தமிழ் மொழிச் செயலாக்கத்தை மேலும் நுண்ணியதாகவும் சவாலானதாகவும் மாற்றுகின்றன. எனவே, அதை பற்றிய அறிவு நமக்கு அவசியமாகிறது

செயற்கை நுண்ணறிவில் தமிழ் எதிர்கொள்ளும் சவால்கள்

தமிழில் இயற்கை மொழிச் செயலாக்கத்தை மேம்படுத்துவதில் பல நடைமுறைச் சவால்கள் உள்ளன. ஆங்கிலம் போன்ற உலகளாவிய மொழிகளுக்குப் பல்வேறு பயன்பாடுகளுக்கான தரவுகள் பரவலாகக் கிடைக்கும் ஆனால் தமிழுக்குப் போதுமான அளவிலான தரவுகள் இன்னும் முழுமையாக உருவாக்கப்படவில்லை எனலாம். மேலும், தமிழ் மொழியில் ஒரு சொல்லுக்குப் பல பொருள்கள் இருப்பதாலும், வாக்கிய அமைப்புகளில்

தமிழ் செயற்கை நுண்ணறிவு ஆய்விதழ்
மதிப்பாய்வு செய்யப்படும் இதழ்
Tamil Journal of Artificial Intelligence
Peer-Reviewed Journal
தொகுதி - 1, இதழ் - 1, ஜனவரி 2026 மின்-ISSN:

பல்வேறு அமைப்புக்கள் காணப்படுவதாலும், கணினிகள் சரியான பொருளைத் துல்லியமாகக் கண்டறிவதில் சிக்கல்கள் எழுகின்றன. இதுதவிர, ஆங்கிலத்திலிருந்து தமிழுக்கோ அல்லது தமிழிலிருந்து பிற மொழிகளுக்கோ மொழிபெயர்க்கும் இயந்திர மொழிபெயர்ப்புக் கருவிகள் முழுமையான துல்லியத்தை அடைய இன்னும் பெருமளவில் மேம்படுத்தப்பட வேண்டிய கட்டாயத்தில் உள்ளன என்பட்டர்ஹெ உண்மை. இத்தகைய சவால்களைக் கடந்து, கணினி இயந்திரம் புரிந்துகொள்ளும் வண்ணம் தமிழை வடிவமைக்கும் இந்தப் பயணமே மொழியின் மறுமலர்ச்சி தொடங்கும் தருணமாக அமைகிறது எனலாம்.

தமிழ் மொழிச் செயலாக்கத்திற்கான ஏ.ஐ கருவிகள்

செயற்கை நுண்ணறிவுத் தொழில்நுட்பத்தின் (AI Technology) அபரிமித வளர்ச்சியால் தற்போது தமிழ் மொழிச் செயலாக்கத்திற்குப் பல்வேறு கருவிகள் உருவாக்கப்பட்டு வருகின்றன. மாணவர்களிடையே கணினிக் கல்வியை ஊக்குவிப்பதற்காக உருவாக்கப்பட்ட 'எழில்' (Ezhil) என்னும் திறந்த மூலத் (Open Access) தமிழ் நிரலாக்க மொழி, தமிழில் மென்பொருள் (Tamil Software) உருவாக்கப் பயிற்சியை வழங்கி வருகிறது. சென்னை ஐ.ஐ.டி (IIT Madras) மற்றும் பிற கல்வி நிறுவனங்களால் உருவாக்கப்பட்ட 'தமிழி என்.எல்.பி' (ThamizhiNLP) திட்டம், உரை சுத்திகரிப்பு (Text Preprocessing), பாகுபடுத்தல் (Parsing), இலக்கணப் பகுப்பு (Part-of-Speech (POS) Tagging (or Syntactic Analysis) மற்றும் உரை மாதிரி உருவாக்கம் (Text Generation (or Language Modeling)) ஆகிய பணிகளைத் தமிழில் செவ்வனே மேற்கொள்கிறது. மேலும், ஏ.ஐ-ஃபார்-பாரத் (AI4Bharat) திட்டத்தின் கீழ் சென்னை ஐ.ஐ.டி வழங்கியுள்ள கருவிகள் மொழிபெயர்ப்பு மற்றும் உரை உருவாக்கத்திற்குப் பெரிதும் உதவுகின்றன எனலாம். குறிப்பாக, இண்டிக்ட்ரான்ஸ் (IndicTrans), இண்டிக்பெர்ட் (IndicBERT) போன்ற செயற்கை நுண்ணறிவு மாதிரிகள் தமிழில் சிறப்பாகச் செயல்படுகின்றன. மிகமுக்கிமாக கூகுள் ட்ரான்ஸ்லேட் (Google Translate) மற்றும் மெட்டா நிறுவனத்தின் 'என்.எல்.எல்.பி' (Meta NLLB - No Language Left Behind) ஆகிய திட்டங்கள் குறைந்த தரவுகளைக் கொண்டு மேம்பட்ட இயந்திர மொழிபெயர்ப்பை வழங்கி வருகின்றன. இதுமட்டுமன்றி, கூகுள் ஸ்பீச் ஏ.பி.ஐ (Google Speech API) மற்றும் மொஸில்லாவின் *காமன் வாய்ஸ்* (Mozilla Common Voice) போன்ற குரல்வழிச் செயற்கை நுண்ணறிவுக் கருவிகள் மூலம் தமிழுக்கான பேச்சு-உரை (Speech-to-Text) மற்றும் உரை-பேச்சு (Text-to-Speech) தொழில்நுட்பங்களும் குறிப்பிடத்தக்க வளர்ச்சியை அடைந்துள்ளன என்பதை காணமுடிகிறது.

பயன்பாடுகளும் எதிர்கால வாய்ப்புகளும்

தற்பொழுது, செயற்கை நுண்ணறிவுக் கருவிகள் (AI Softwares) கல்வி மற்றும் ஆராய்ச்சித் துறைகளிலும் பெரும் தாக்கத்தை ஏற்படுத்தி வருகின்றன. உரை சுருக்கம் (Text Summarization or Automatic Text Summarization), வினா-விடை உருவாக்கம் (Question-Answer Generation or Automated Question Generation, QAG) மற்றும் மாணவர் செயல்திறன் மதிப்பீடு (Student Performance Evaluation or Automated Assessment / Learning Analytics) போன்றவை இத்தொழில்நுட்பத்தின் மூலம் எளிதாக மேற்கொள்ளப்பட்டு வருகின்றன. சங்க இலக்கியங்கள் உள்ளிட்ட பண்டைய தமிழ் நூல்களை எண்ணிம (Digital) வடிவில் ஆவணப்படுத்தவும், அவற்றை ஆழமாக ஆய்வு (Deep Research) செய்யவும் இக்கருவிகள் உதவுகின்றன. எதிர்காலத்தில், பல்வேறு தமிழ்ப் பேச்சு வழக்குகளை (Dialects) அடிப்படையாகக் கொண்ட பன்முக உதவிகள், செயற்கை நுண்ணறிவு வழியிலான தமிழ் மருத்துவ ஆலோசனைகள், விவசாயம் சார்ந்த வழிகாட்டுதல்கள் மற்றும் இலக்கிய விமர்சனங்களுக்கான கணினி மதிப்பீடுகள் எனப் பல துறைகளிலும் செயற்கை நுண்ணறிவுத்

தொகுதி - 1, இதழ் - 1, ஜனவரி 2026 மின்-ISSN:

தொழில்நுட்பம் தமிழ் மொழியோடு இணைந்து செயல்படும் வாய்ப்புகள் உள்ளன எனலாம். அதற்கான நிலையான வழிகாட்டு முறைகளை (SOP) உருவாக்குவது மிக முக்கியமாகும்.

சமூகவியல் மற்றும் நெறிமுறைச் சிந்தனைகள்

இவை அனைத்திற்கும் மேலாக, தமிழுக்கான செயற்கை நுண்ணறிவுக் கருவிகள் உருவாக்கப்படும்போது சமூகவியல் மற்றும் நெறிமுறைச் சிந்தனைகளைக் கருத்தில் கொள்வது மிகவும் முக்கியம். தரவுகளில் நம்பகத்தன்மையை உறுதி செய்தல், மொழி மரபைக் காத்தல், சமூகப் பாகுபாடுகளைத் தவிர்த்தல் மற்றும் பண்பாட்டு அடையாளத்தை நிலைநிறுத்துதல் ஆகிய பண்புகளோடு இத்தொழில்நுட்பம் வளர வேண்டும். ஒருதலைப்பட்சமான அல்லது தவறான தரவுகள் சமூகத்தில் எதிர்மறையான தாக்கங்களை ஏற்படுத்தக்கூடும். அதனால், இத்தகைய விழிப்புணர்வு இல்லையெனில், தொழில்நுட்ப உலகில் தமிழ் வெறும் 'பயனர் மொழியாக' மட்டுமே விளங்கும் அபாயம் உள்ளது. அதனை தவிர்க்க இம்மாதிரியான முறைகளை அறிந்து கொள்வது முக்கியமாகும். வருங்காலங்களில், பள்ளி மற்றும் கல்லூரியிலும் AI தொழில்நுட்ப குறியீடுகளை பாடமாக இணைப்பது அவசியமாகின்றது.

முடிவுரை

செயற்கை நுண்ணறிவு (AI) மற்றும் இயற்கை மொழிச் செயலாக்கம் (NLP) ஆகியவை தமிழ் மொழிக்காகப் பல புதிய வாய்ப்புகளை உருவாக்கி வருகின்றன. சரியான தரவுகளையும் சமூகச் சிந்தனையையும் உள்ளடக்கி உருவாக்கப்படும் செயற்கை நுண்ணறிவுக் கருவிகள், தமிழ் மொழியை உலகளாவிய மையத்தில் நிலைநிறுத்துவதோடு மட்டுமல்லாமல், அடுத்த தலைமுறைக்கு நமது பண்பாட்டு அடையாளத்தைக் கடத்தும் பாலமாகவும் அமையும். இன்றைய நிலையில் தமிழில் செயற்கை நுண்ணறிவு வெறும் தொழில்நுட்ப முன்னேற்றமாக மட்டுமல்லாமல் தமிழ் மொழி கணினி பயன்பாட்டில் சிறக்க ஓர் அரியதோர் வாய்ப்பாக அமையும். மேலும், தமிழ் மொழியை கணினி அறிவியல் சார்ந்து ஆய்வு செய்து படிக்க இம்மாதிரியான பகுப்பு முறைகள் உதவும்.

குறிப்புகள்

[1] ஆர்டிலா, ஆர்., பிரான்சன், எம்., டேவிஸ், கே., ஹென்ரெட்டி, எம்., கோலர், எம்., மேயர், ஜே., மோரேஸ், ஆர்., சாண்டர்ஸ், எல்., டயர்ஸ், எஃப். எம்., & வெபர், ஜி. (Ardila, R. et al.). (2020). காமன் வாய்ஸ்: ஒரு மாபெரும் பன்மொழிப் பேச்சுத் தரவுத்தொகுப்பு (Common voice: A massively-multilingual speech corpus). 12-வது மொழி வளங்கள் மற்றும் மதிப்பீட்டு மாநாட்டுக் கோவை (Proceedings of the 12th Language Resources and Evaluation Conference), 4218–4222. https://aclanthology.org/2020.lrec-1.520

[2] கக்வானி, டி., குஞ்சுகுட்டன், ஏ., கொல்லா, எஸ்., கோகுல், என். சி., பட்டாச்சார்யா, பி., காப்ரா, எம். எம்., & குமார், பி. (Kakwani, D. et al.). (2020). இண்டிக் என்.எல்.பி சூட் (IndicNLPSuite): இந்திய மொழிகளுக்கான ஒருமொழித் தரவுத்தொகுப்புகள், மதிப்பீட்டு அளவுகோல்கள் மற்றும் முன்-பயிற்சி பெற்ற பன்மொழி மாதிரிகள். கணக்கீட்டு மொழியியல் சங்கத்தின் கண்டுபிடிப்புகள் (Findings of the Association for Computational Linguistics: EMNLP 2020), 4948-4961. https://doi.org/10.18653/v1/2020.findings-emnlp.445

[3] கூகுள் (Google). (2024). கூகுள் கிளவுட் ஸ்பீச்-டு-டெக்ஸ்ட் ஏ.பி.ஐ (Google Cloud Speech-to-Text API) [கணினி மென்பொருள்]. கூகுள் கிளவுட். https://cloud.google.com/speech-to-text

[4] கூகுள் (Google). (2024). கூகுள் கிளவுட் ட்ரான்ஸ்லேஷன் ஏ.பி.ஐ (Google Cloud Translation API) [கணினி மென்பொருள்]. கூகுள் கிளவுட். https://cloud.google.com/translate

தொகுதி - 1, இதழ் - 1, ஜனவரி 2026 மின்-ISSN:

[5] சர்வேஸ்வரன், க., & டயஸ், ஜி. (Sarveswaran, K., & Dias, G.). (2020). தமிழி என்.எல்.பி (ThamizhiNLP): தமிழ் இயற்கை மொழிச் செயலாக்கத்திற்கான ஒரு கட்டமைப்பு. 12-வது மொழி வளங்கள் மற்றும் மதிப்பீட்டு மாநாட்டுக் கோவை (Proceedings of the 12th Language Resources and Evaluation Conference), 6745–6753. https://aclanthology.org/2020.lrec-1.833

[6] என்.எல்.எல்.பி குழு (NLLB Team), கோஸ்டா-ஜுஸ்ஸா, எம். ஆர்., கிராஸ், ஜே., மற்றும் பலர். (2022). எந்த மொழியும் பின்தங்கவில்லை: மனிதர்களை மையமாகக் கொண்ட இயந்திர மொழிபெயர்ப்பை அளவிடுதல் (No language left behind: Scaling human-centered machine translation). arXiv. https://doi.org/10.48550/arXiv.2207.04672

[7] முத்தையா, மு. (Muthiah, M.). (2023). எழில்: ஒரு தமிழ் நிரலாக்க மொழி (Ezhil: A Tamil programming language) (பதிப்பு 2.0) [கணினி மென்பொருள்]. எழில் மொழித் திட்டம் (Ezhil Language Project). https://ezhillang.org/

நிதிசார் கட்டுரையாளர் உறுதிமொழி: இல்லை
கட்டுரையாளர் நன்றியுரை: இல்லை
கட்டுரையாளர் உறுதிமொழி: இக்கட்டுரையில் எவ்வித முரண்பாடும் இல்லை என்று உறுதிமொழி அளிக்கிறேன்.