NASSCOM®

Ai Gamechangers

AI GAMECHANGERS: ACCELERATING INDIA WITH INNOVATION

COMPENDIUM OF 50 AI INNOVATION STORIES

2021 EDITION

INNOVATION PARTNER

Microsoft

KNOWLEDGE PARTNER

Deloitte.

Natural Language Processing
GAMECHANGER

IIT Madras: Bridging the language divide

Samanantar, described as translation corpora and tools for the next billion users, is a vernacular translation system that can translate legal documents in the judicial domain.

Ages ago, following the great flood, some people decided to migrate eastward to build a city and a tower at Babel. The tower was meant to be the tallest one ever so that it would reach the heavens, bringing the men who built it eternal fame and glory. Seeing this, God − in an attempt to divide the human race − created many new languages, to confound their speech, and scattered them around the world.

The famous Tower of Babel story from the ancient scriptures might be a myth. But strong barriers that exist in our civilisation as a result of the linguistic divide is a reality − even in the modern age of digital and emerging technologies. And nowhere this is more visible than in India − a land of vast linguistic diversity with over 500 languages.

Since its independence, India has been facing its own 'Tower of Babel' moments in its governance, especially in the judiciary. The country's judiciary has a hierarchal system that starts from the Munsif Courts and Sessions Courts at the local level, to high courts at the state level, and finally the Supreme Court at the national level. In most cases, lower tier courts use local or regional languages, while the Supreme Court and many high courts use English in their proceedings as specified in Article 348 of the Constitution.

As a result, many of the proceedings from the Supreme Court is often inaccessible to non-English speakers of the country, hampering the fundamental right to have equality before the law.

In the 2018-19 Annual Report, the Supreme Court has declared its commitment to translating judgments into the vernacular languages of litigants. Recently, it has started translating its proceedings to nine of the 22 scheduled languages. However, this has been done manually, which is time-consuming, expensive, and leads to more delays in the already snail-paced judiciary process.

On the bright side, major advancements have been happening in an AI sub-branch called Natural Language Processing or NLP. Natural Language Processing makes machines process, understand and generate human languages. And only an Natural Language Processing-powered automatic translation system can provide the solution to the language problems in the judicial system. However, that requires an accurate, free, open-source, automatic translation system that can process Indian languages, which is almost non-existent. And that is where Samanantar - an AI language translation model built by the IIT Madras faculty Mitesh Khapra and Pratyush Kumar, becomes a Gamechanger.

The 'Samanantara Yatra'

Samanantar's origin goes back to the days when Mitesh who has been working in the space of Natural Language Processing as part of his PhD, and Pratyush who was focusing on the systems side of AI, came together to create the AI4Bharat initiative.

> We understood that in today's world of deep learning, we need to combine domain expertise with the technology to make any impact on society, and we chose the domain of Natural Language Processing.
> We recognised that we need to combine Natural Language Processing with systems thinking to build large systems, to collect large amounts of data, and then address some of these problems in very standardised ways.
>
> **Pratyush Kumar**
> Associate Professor, IIT Madras

Samanantar, described as translation corpora and tools for the next billion users, can translate legal documents in the judicial domain. It is currently deployed in the Supreme Court of India as a pilot project.

We felt that before starting with translation, sentiment detection and question answering, it is good to start with foundational blocks, which is to simply create a large corporate of Indian language text. And eventually we were able to increase the size of corpus available by an order of magnitude of 10 in many Indian languages. - Pratyush Kumar

Samanantar's journey had two critical components − the Indic language corpus created for training models and the Vector Space Models used for machine translation. The first one was the biggest hurdles to overcome in India's vernacular Natural Language Processing space due to lack of training data for Indic languages.

In fact, one of the key foundations of Samanantar is the 46 million parallel sentences collected by the team using smart tools from the web on which the models were trained.

The second core component of Samanantar is the vector space models that were then trained on this rich Indic language corpus.

Vector space models are algebraic models that are often used to represent text as a vector of identifiers. With these models, one can identify whether various texts are similar in meaning, regardless of whether they share the same words. The team then used efficient approximate nearest neighbour search using FAISS to search over 100 million sentences to find a matching sentence resulting in translation.

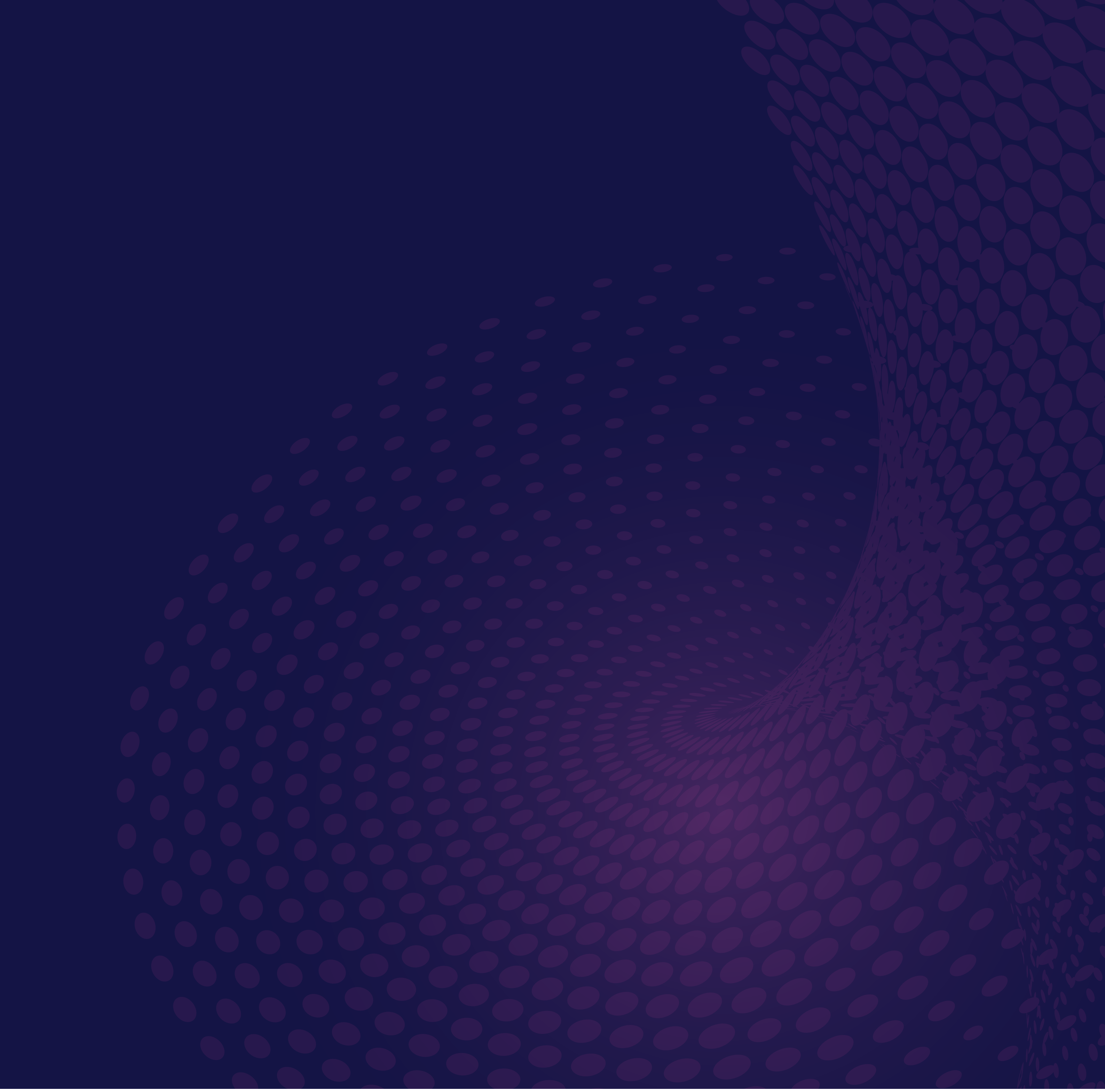
Presently, Samanantar offers English to Indic and Indic to English translation and supports 11 Indian languages. Apart from the Supreme Court of India, Samanantar is now being used by Bangladesh Supreme Court, C-DAC, Pratham Books, and Ek Step Foundation for translation works. It has been known to provide an accuracy rate much better than similar solutions from tech giants, such as Google and Microsoft.

According to Pratyush, the team is also working on an Indic to Indic translation model with languages Tamil and Oriya.

Indic languages face a unique challenge in their continuity and relevance, in the age of the Internet – one that predominantly caters to English. Efforts made by companies such as Samanantar are moving in the direction of making the internet more equitable and suited to the preferences of a land characterised by multiple languages.

The AI model we used can take a Hindi sentence and a sentence in another language such as English and map it to a common vector space where it can find parallel sentences based on distance - Mitesh Khapra
IIT Madras faculty

> The goal here is to bring parity in AI technology for Indian languages with English and we want to build these solutions for this long tail of Natural Language Processing tasks for as many Indian languages as possible.
>
> **Mitesh Khapra**
> Associate Professor, IIT Madras



NASSCOM®

www.nasscom.in

research@nasscom.in

+91-120-4990111

www.community.nasscom.in

NASSCOM Plot 7 to 10, Sector 126,
Noida - 201303

/NASSCOMOfficial

/nasscom

/NasscomVideos