DE-III
Natural Language Processing

Dr. Yashodhara Haribhakta

Department of Computer Engg. & I.T.,
College of Engineering Pune

Email: *ybl.comp@coep.ac.in*

Theory Assessment

T1/Quiz/Surprise test – 20 marks

T2/Quiz/Surprise test – 20 marks

End-Semester – 60 marks

Note :

1) Minimum marks – 40 marks for passing in this subject.

NLP-Syllabus

NLP-Syllabus

(CT)Natural Language Processing

Teaching Scheme:
Lectures : 3 Hrs/week

Examination Scheme:
Continuous evaluation – 100 marks
Assignment/Quizzes – 40 marks
End Sem Exam - 60 marks

Course Outcomes:

Students will be able to:

1. Understand basic text processing techniques in NLP.
2. Analyse morphological analyzers and stemmers.
3. Build language models.
4. Design, implement and evaluate part-of-speech taggers and parsers.
5. Understand knowledge based wordnet and apply it for Word SenseDisambiguation.

Unit I : Introduction: Introduction , motivation, word tokenization, word normalization, word level morphology- morphological analysis and synthesis, stemming - porters algorithm, levenshtein distance measure **[6 Hrs]**

Unit II: POS Tagging: Sequence labeling tasks of NLP, POS tagging, POS tagsets, Hidden Markov Model, Viterbi algorithm, Baum Welch Algorithm . **[6 Hrs]**

Unit III: Language Modeling: Introduction to N-gram, probability estimation for n-gram, evaluation and perplexity, smoothing techniques, Named-Entity recognition. **[6 Hrs]**

NLP-Syllabus

Unit IV: Parsers: Constituency and Dependency parsers, Constituency parser -Syntactic structure, parsing methodology, different parsing algorithms, parsing in case of ambiguity; probabilistic parsing , the CKY algorithm, issues in parsing, Dependency parsing- Syntactic structure, parsing methodology, Transition-Based Dependency Parsing , Graph-Based Dependency Parsing, Evaluation, co-reference resolution, Named-entity recognition. **[8 Hrs]**

Unit V : WordNet: Word Senses, word relations, word similarity and thesaurus methods, Word sense disambiguation, Knowledge base and supervised WSD, WordNet , Unsupervised based WSD. **[6 Hrs]**

Unit VI : Applications of NLP: Question/Answering System, Text Summarization, Sentiment Analysis, Information extraction **[4 Hrs]**

Text Books:

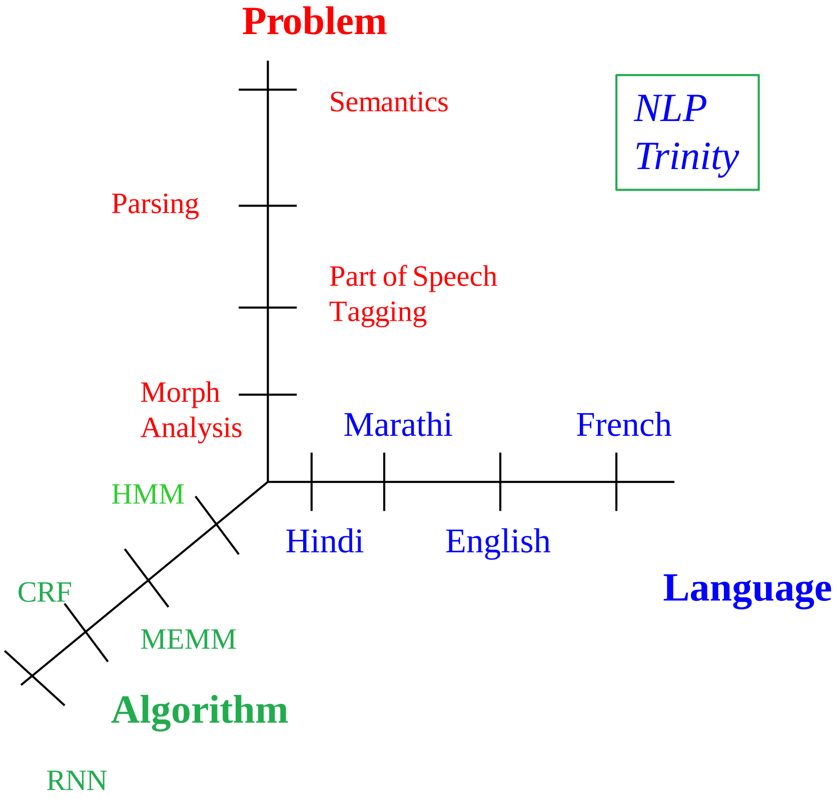
- Daniel Jurafsky and James H. Martin, "Speech and Language Processing", Second Edition, Prentice Hall, 2008.
- Allen James, "Natural Language Understanding", Second Edition, Benjamin/Cumming, 1995.
- Chris Manning and HinrichSchuetze, "Foundations of Statistical Natural Language Processing", MIT Press.

Reference Books:

- Journals : Computational Linguistics, Natural Language Engineering, Machine Learning, Machine Translation, Artificial Intelligence .

Course Objectives

- Introduce the fundamental concepts and techniques of natural language processing (NLP) by studying the phonological, morphological, syntactic and semantic processing.
- To gain an in-depth understanding of algorithms available for the processing of linguistic text information and the underlying computational properties of natural languages .



Why do we need to study NLP?

Introduction

- **What is NLP?**

Introduction

- **What is NLP?**
  - Processing text data so that able to infer some information which is useful.

Introduction

- **What is NLP?**
  - Processing text data so that able to infer some information which is useful.
- **What is the main goal of NLP?**

Introduction

- **What is NLP?**
  - Processing text data so that able to infer some information which is useful.
- **What is the main goal of NLP?**

  1. **Fundamental and Scientific Goal**
     - ➢ Deep Understanding of natural language

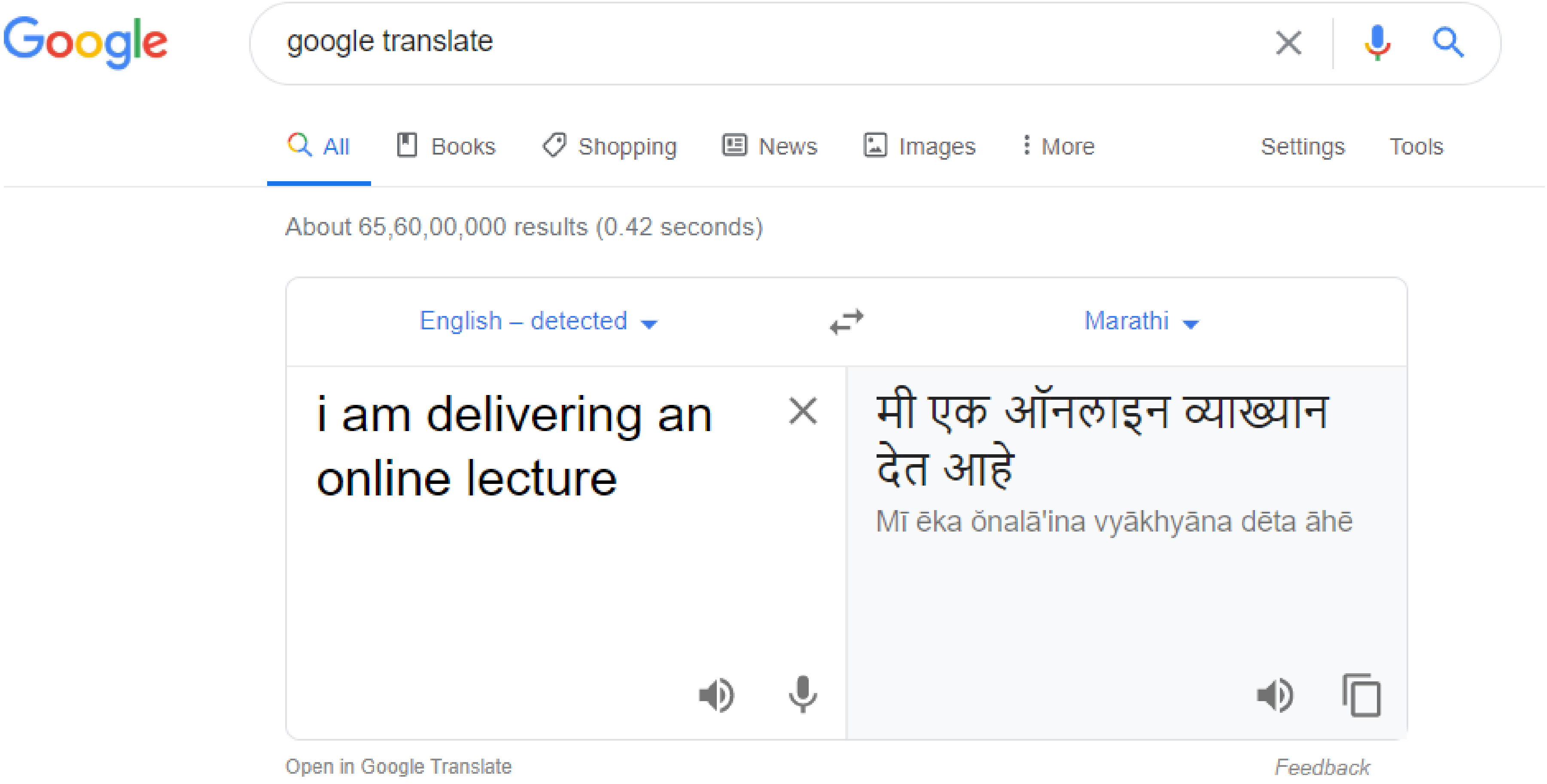
Introduction

- **What is NLP?**
  - Processing text data so that able to infer some information which is useful.
- **What is the main goal of NLP?**

  1. **Fundamental and Scientific Goal**
     - ➢ Deep Understanding of natural language
  2. **Practical and Engineering goal**
     - ➢ Design, implement and test systems that process natural language for practical applications

Engineering Goals: Some examples

Language Translation



google translate

About 65,60,00,000 results (0.42 seconds)

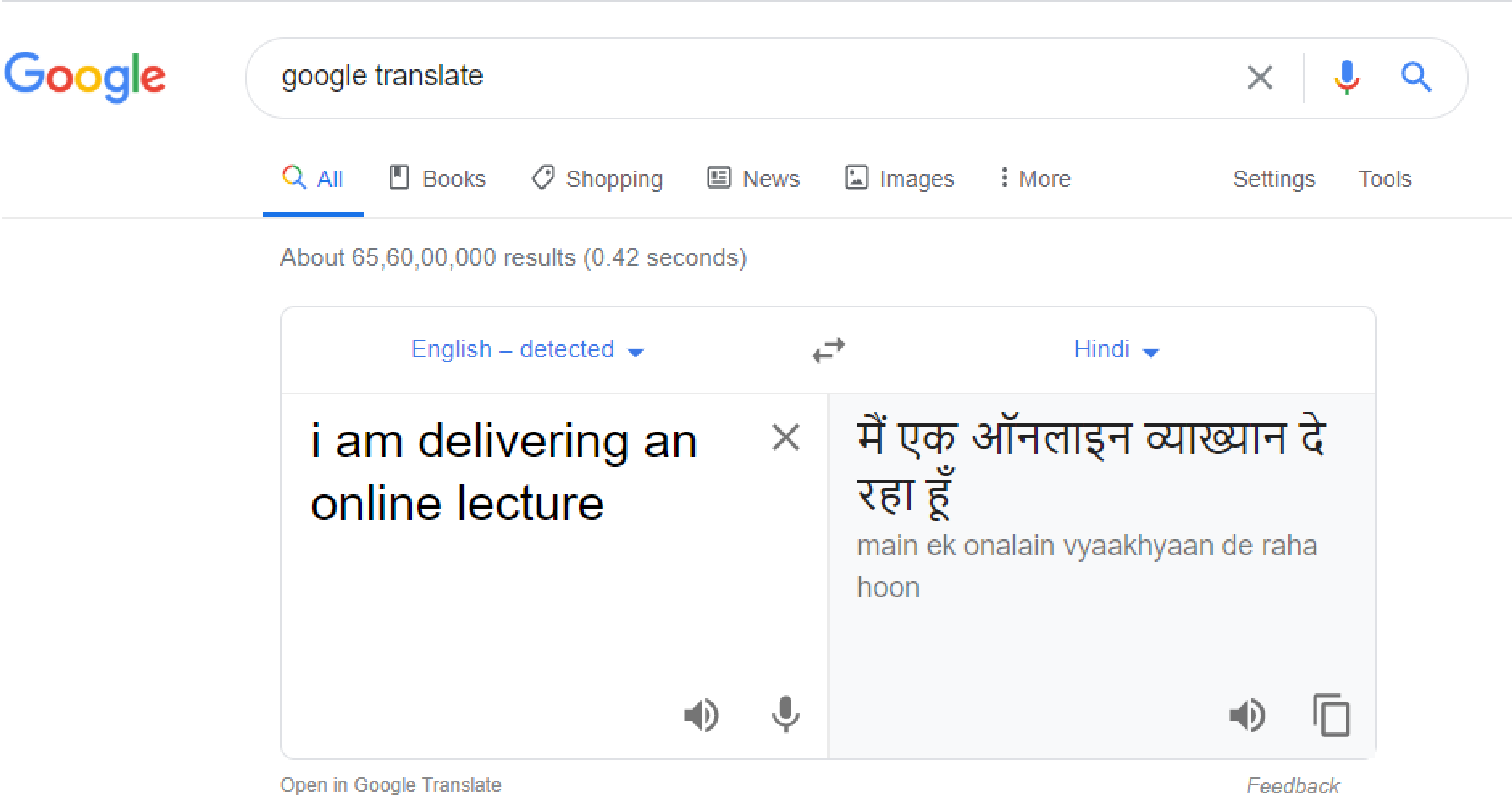
English – detected → Marathi

i am delivering an online lecture

मी एक ऑनलाइन व्याख्यान देत आहे

Mī ēka ŏnalā'ina vyākhyāna dēta āhē

Language Translation



google translate

About 65,60,00,000 results (0.42 seconds)

English – detected → Hindi

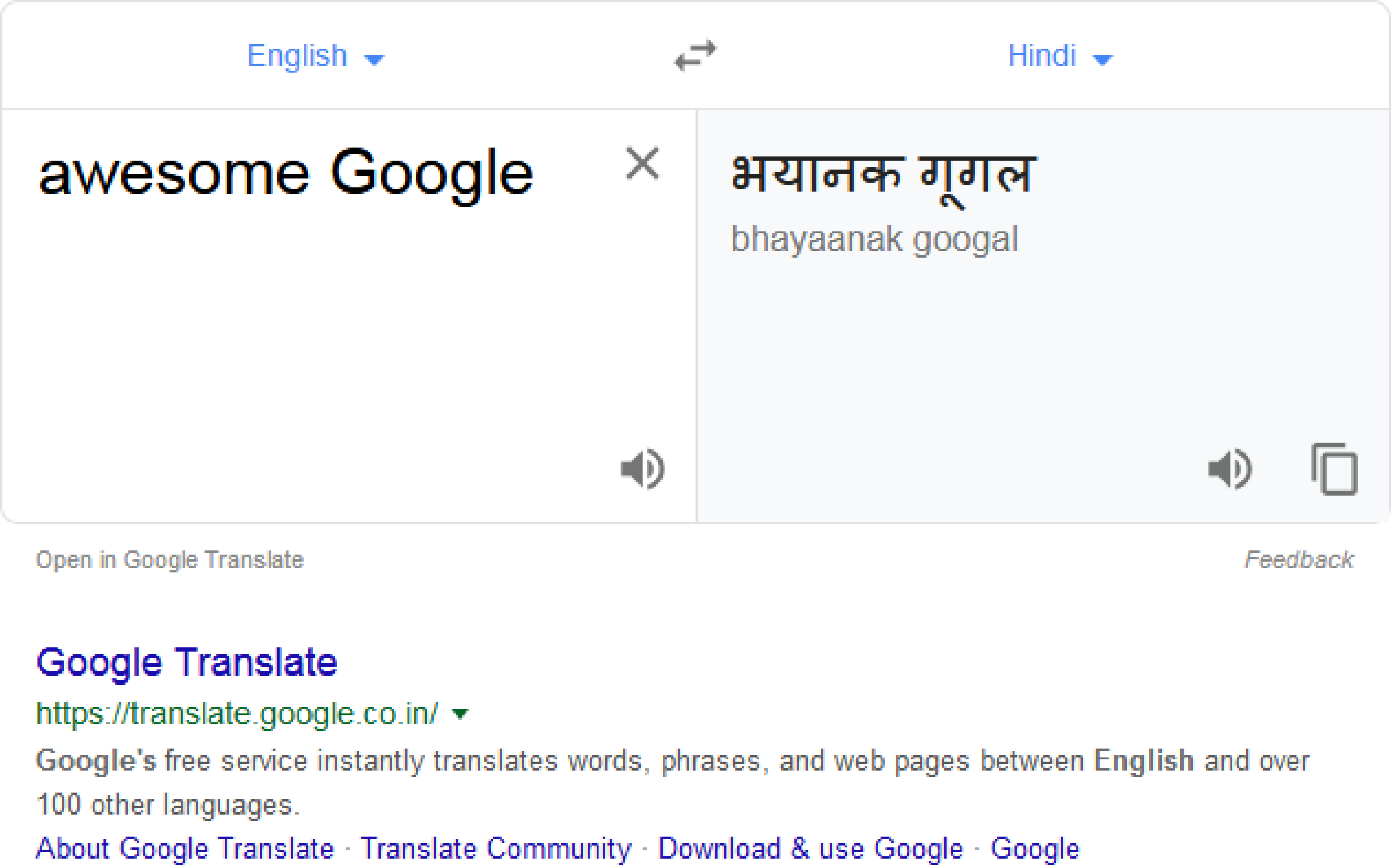
i am delivering an online lecture

मैं एक ऑनलाइन व्याख्यान दे रहा हूँ

main ek onalain vyaakhyaan de raha hoon

Language Translation

One year back



awesome Google

भयानक गूगल

bhayaanak googal

Google Translate

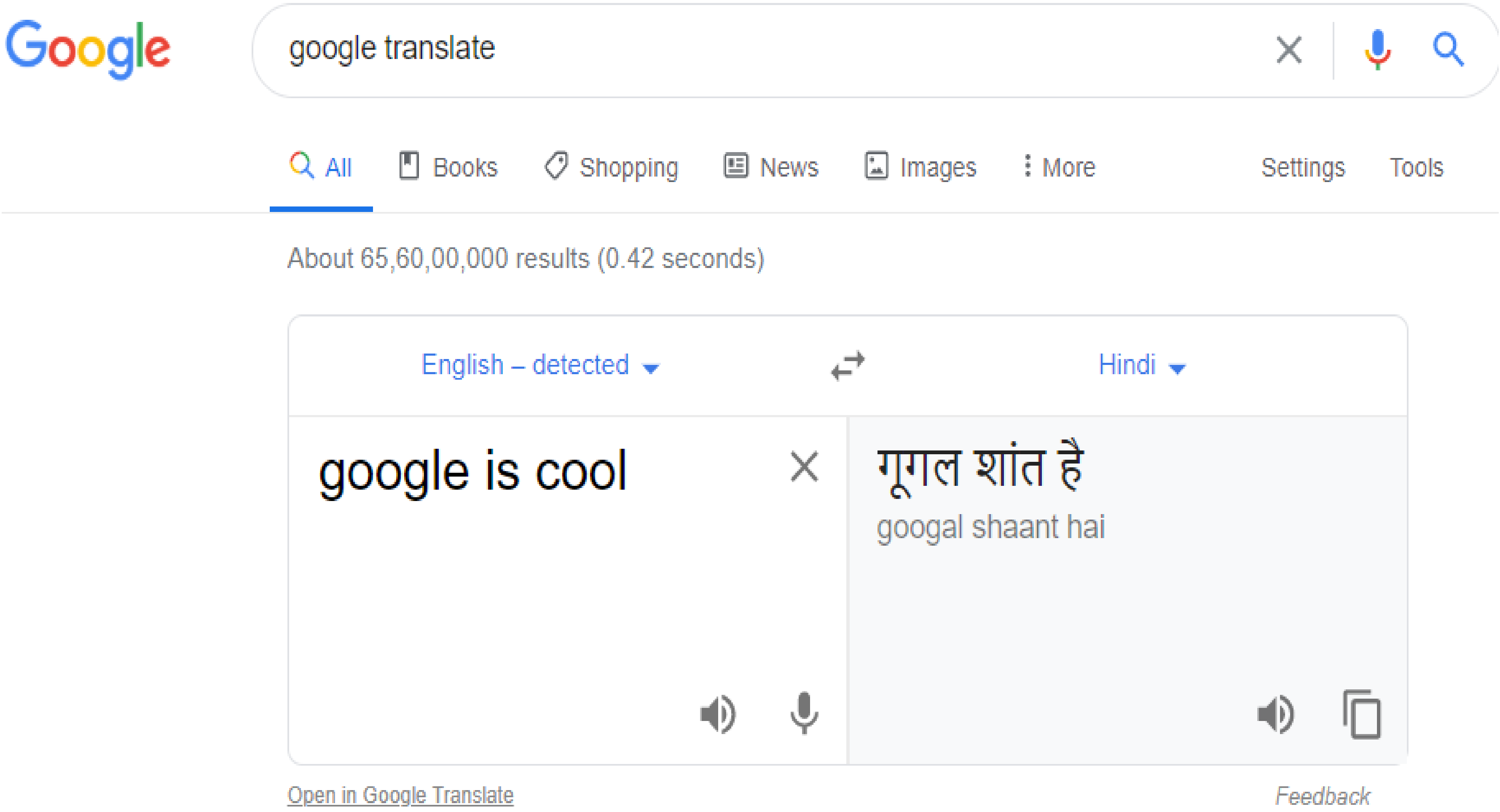
https://translate.google.co.in/

Google's free service instantly translates words, phrases, and web pages between **English** and over 100 other languages.

About Google Translate · Translate Community · Download & use Google · Google

Language Translation

One year back



google translate

About 65,60,00,000 results (0.42 seconds)

English – detected

Hindi

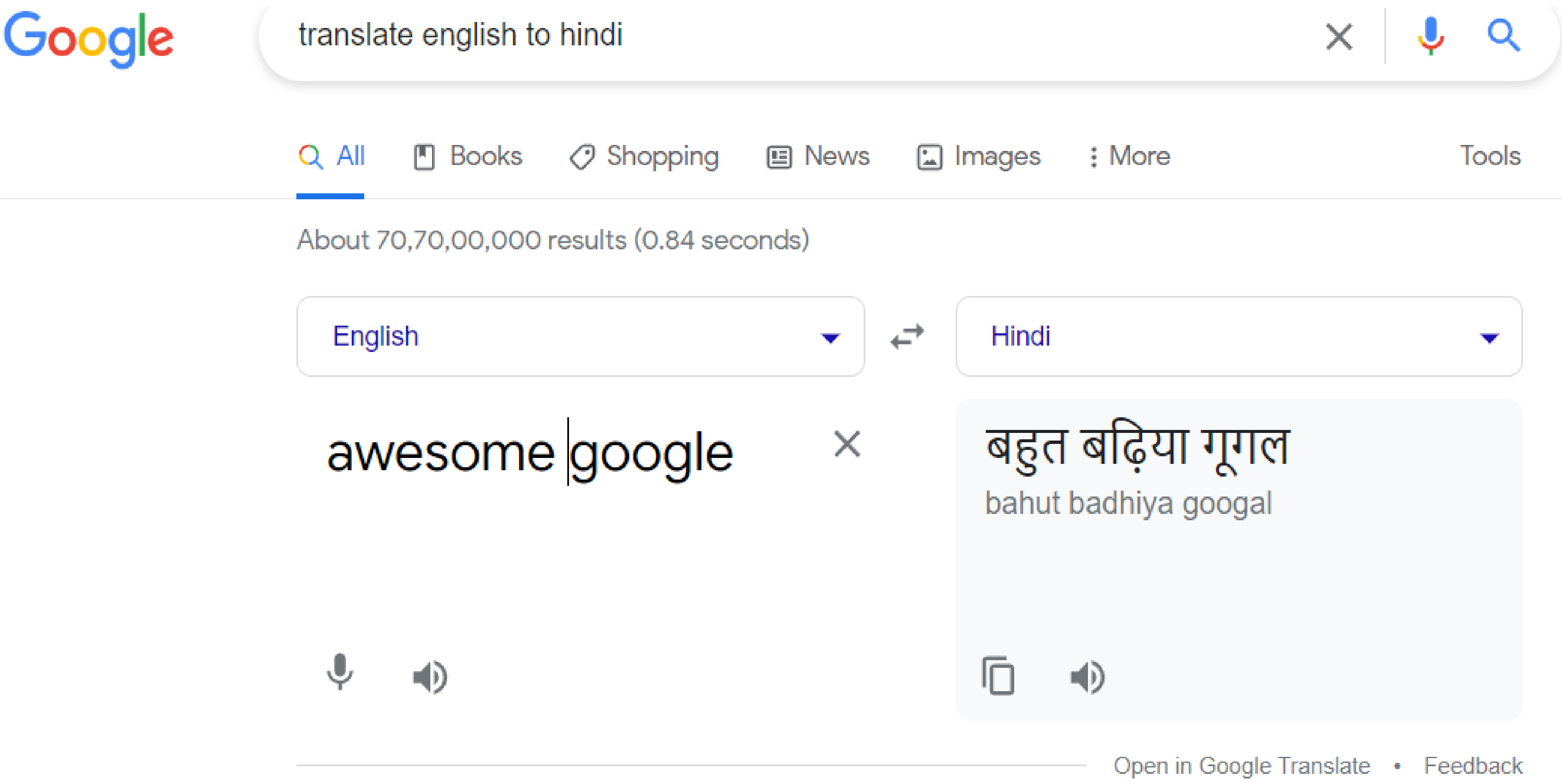
google is cool

गूगल शांत है

googal shaant hai

Language Translation

Today’s result



translate english to hindi

About 70,70,00,000 results (0.84 seconds)

English → Hindi

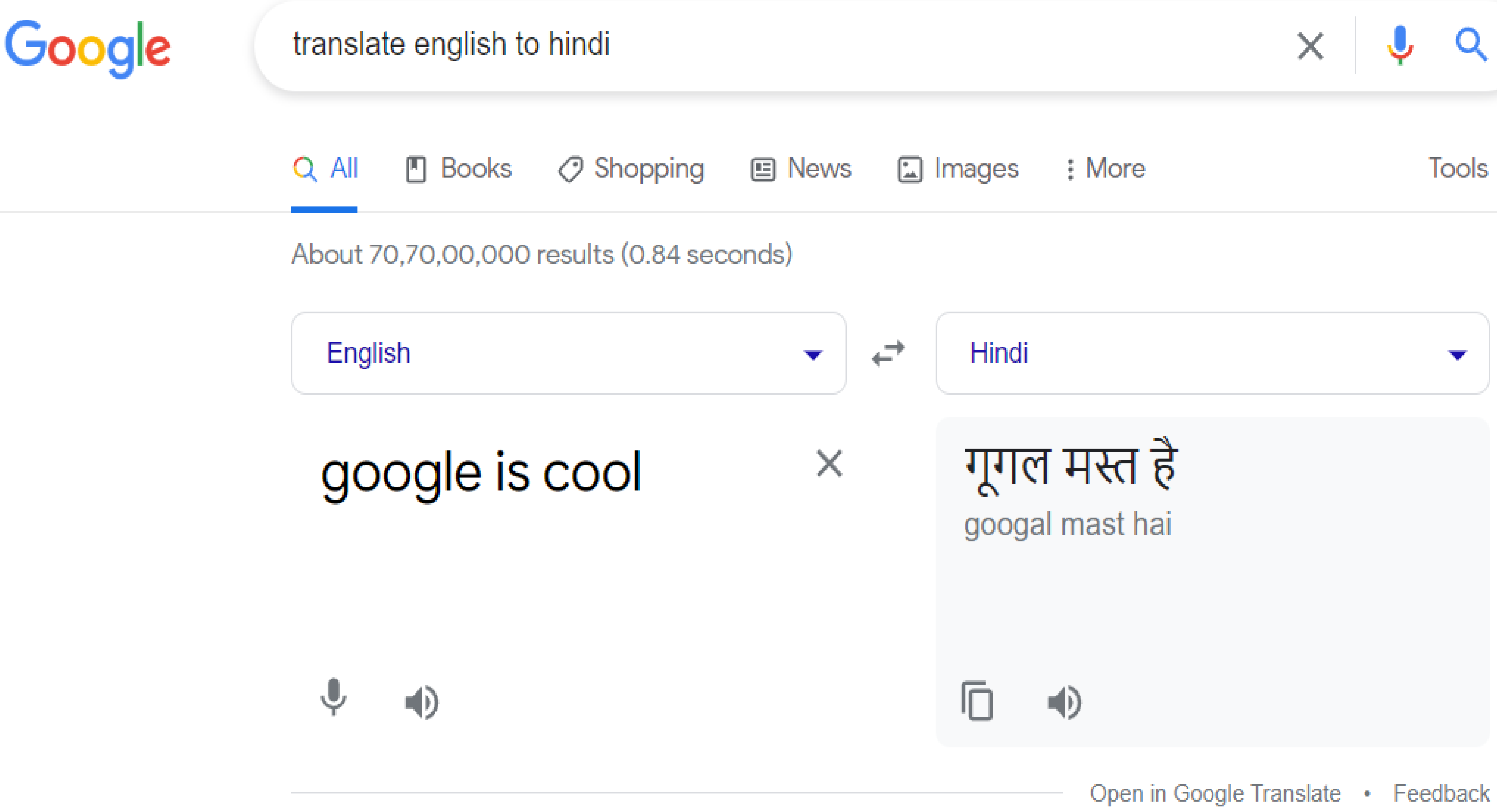
awesome google

बहुत बढ़िया गूगल

bahut badhiya googal

Language Translation

Today’s result



translate english to hindi

About 70,70,00,000 results (0.84 seconds)

English → Hindi

google is cool

गूगल मस्त है

googal mast hai

Applications

1. Word correction (Auto correction)
2. Query Completion (Auto Completion)
3. Question Answering systems (Google BERT)
4. Code to document generation(Microsoft codeBERT)
5. ChatGPT (openAI)
6. Sentiment Analysis(roBERT, BERT, Distilbert)
7. Text summarization (T5, GPT2, GPT3 , XLNet)
8. Named Entity Recognition (NER-BERT)
9. Targeted Marketing….

Course Outcomes

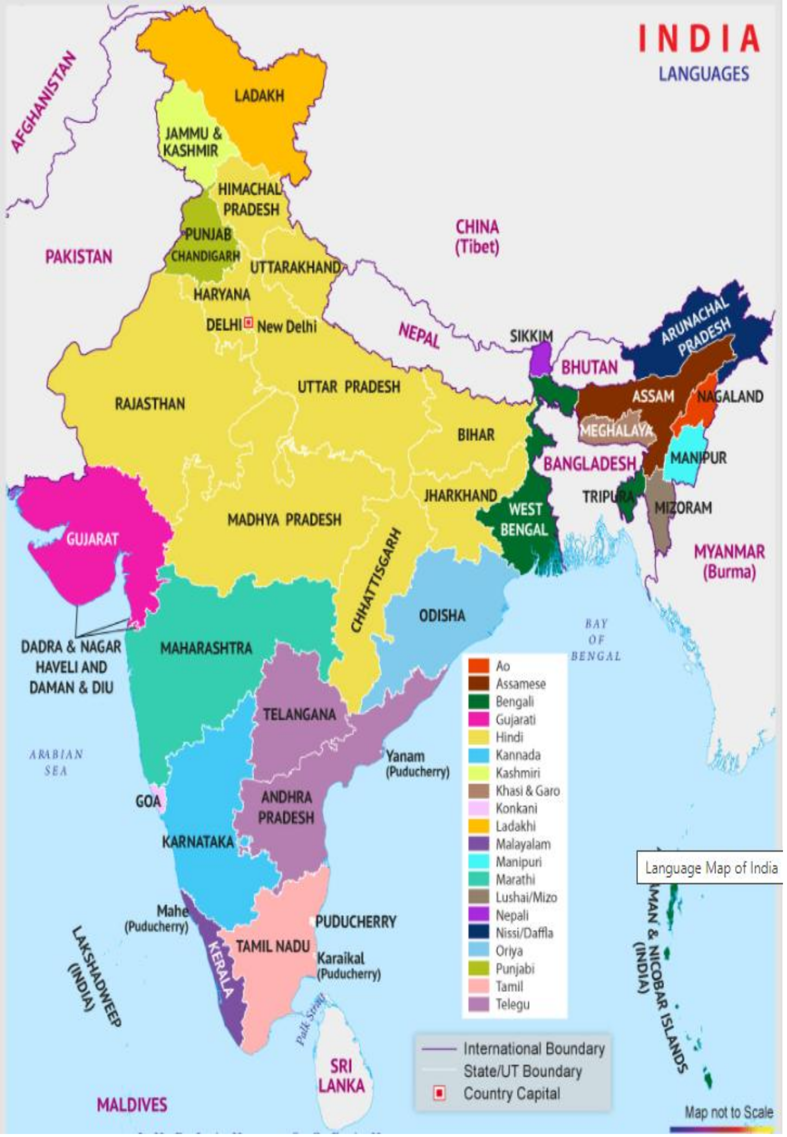
Students should be able to:

1. Demonstrate the understanding of basic text processing techniques in NLP
2. Analyze the morphological analyzers and stemmers
3. Build language models and demonstrate WSD using knowledge base WordNet for English langauge
4. Design, Implement and evaluate the POS taggers and parsers

Natural Language Processing: Background & Relevance in Indian Scenario

Multilinguality: Indian situation

- Language families
  - Indo Aryan
  - Dravidian
  - Austro-Asiatic
  - Tibeto-Burman
- Languages that are ranked within 20 in the world in terms of the populations speaking them, are
  - Hindi : 3rd (~350 milion)
  - Bangla: 7th (~230 million)
  - Marathi:15th (~84 million)



Background: Indian Context

- India is a multi-lingual country with great linguistic and cultural diversities
- 22 official languages mentioned in the Indian constitution
- However, Census of India in 2011 reported-
  - **121 major languages**
  - **1,599 other regional languages**
  - **2,371 scripts**
  - **30 languages** are spoken by more than **one million native speakers**
  - **121** are spoken by more than **10,000 people**
- **20%** understand English
- **80%** cannot understand

https://en.wikipedia.org/wiki/Languages_of_India

Background

- Phenomenal growth in the number of internet users, social media (***Facebook,Twitter*** etc.)
- Increasing tendency of using Indian language contents for exchanging information
- **Digital divide** cannot be tackled unless citizens are given flexibility in **communicating in their own languages**

Natural Language Processing (NLP) that deals with developing theories and techniques for effective communication in human languages play an important role towards creating this digital society

Motivation

TDIL: MeiTY, Govt. of India

TDIL : Technology Development for Indian Languages Programme initiated by the Ministry of Electronics & Information Technology, Govt. of India

Objective:

- objective of developing Information Processing Tools and Techniques to facilitate human-machine interaction without language barrier;
- creating and accessing multilingual knowledge resources; and
- integrating them to develop innovative user products and services.

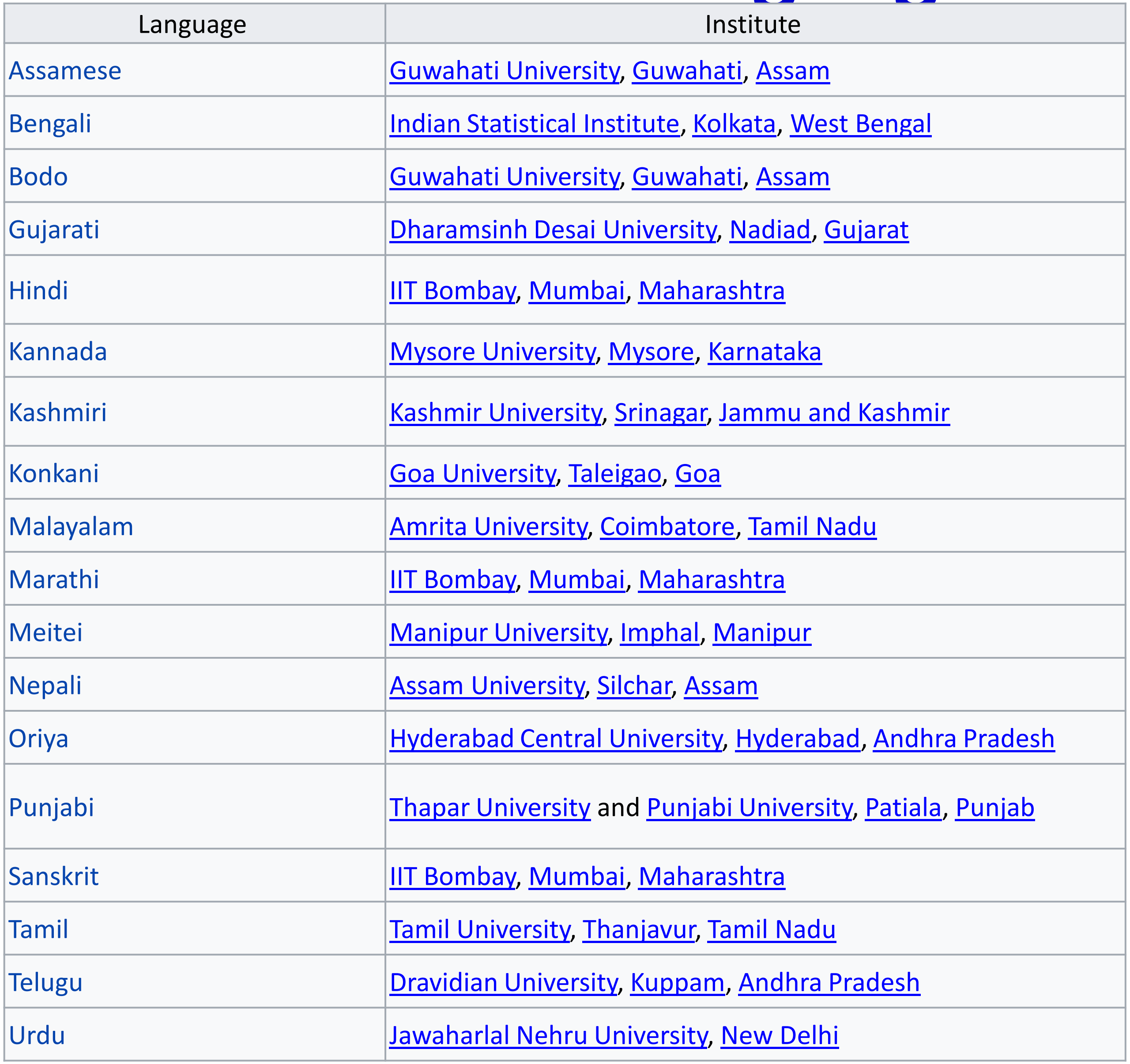
TDIL: Some major Machine Translation Projects

1. **Development of English to Indian Language Machine Translation System** (**Anuvadaksh**): Translator for English to Hindi/ Marathi/ Bangla/ Oriya/ Tamil/ Urdu/ Gujrati/ Bodo
2. **Development of English to Indian Language Machine Translation System with Angla-Bharti Technology**: ANGLABHARTI represents a machine-aided translation methodology specifically designed for translating English to Indian languages, like, English to Bangla/ Punjabi/ Malaylam/ Urdu/ Hindi/ Telugu
3. **Development of Indian Language to Indian Language Machine Translation System (Sampark)**- 18 pairs of languages, like, -Hindi to Bengali, Bengali to Hindi, Marathi to Hindi, Hindi to Marathi, Hindi to Punjabi, Punjabi to Hindi, Hindi to Tamil, Tamil to Hindi, Hindi to Kannada, Kannada to Hindi, Hindi to Telugu, Telugu to Hindi, Hindi to Urdu, Urdu-Hindi, Malaylam to Tamil,Tamil to Malaylam,Tamil to Telugu, Telugu to Tamil

TDIL: Some major initiatives

- Development of Cross-Lingual Information Access (CLIA)
  - Assamese, Bengali, Hindi, Oriya, Punjabi, Tamil, Telugu, Marathi, Gujarati
- Development of Robust Document Analysis & Recognition System for Indian Languages (OCR) - 14 languages
  - Assamese, Bengali, Devanagri, Gujarati, Gurumukhi, Kannada, Malaylam, Manipuri, Marathi, Oriya, Tamil, Telugu, Tibetan, Urdu
- Development of Text to Speech System in Indian Languages
- Development of Automatic Speech Recognition System in Indian Languages
- Development of Sanskrit Machine Translation System
- ***Development of Hindi to English Machine Translation in Judicial Domain***

Languages and the Institutes working on different language



Language	Institute
Assamese	Guwahati University, Guwahati, Assam
Bengali	Indian Statistical Institute, Kolkata, West Bengal
Bodo	Guwahati University, Guwahati, Assam
Gujarati	Dharamsinh Desai University, Nadiad, Gujarat
Hindi	IIT Bombay, Mumbai, Maharashtra
Kannada	Mysore University, Mysore, Karnataka
Kashmiri	Kashmir University, Srinagar, Jammu and Kashmir
Konkani	Goa University, Taleigao, Goa
Malayalam	Amrita University, Coimbatore, Tamil Nadu
Marathi	IIT Bombay, Mumbai, Maharashtra
Meitei	Manipur University, Imphal, Manipur
Nepali	Assam University, Silchar, Assam
Oriya	Hyderabad Central University, Hyderabad, Andhra Pradesh
Punjabi	Thapar University and Punjabi University, Patiala, Punjab
Sanskrit	IIT Bombay, Mumbai, Maharashtra
Tamil	Tamil University, Thanjavur, Tamil Nadu
Telugu	Dravidian University, Kuppam, Andhra Pradesh
Urdu	Jawaharlal Nehru University, New Delhi

Thanks