Challenges of Processing South Asian Languages (CPSAL)

Kengatharaiyer **Sarves**waran
University of Konstanz, Germany
University of Jaffna, Sri Lanka.

Tafseer Ahmed
Senior NLP Scientist
Alexa Translations

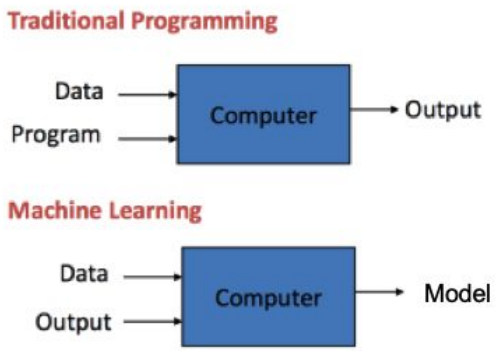
ESSLLI 2024 European Summer School of Logic, Language and Information Leuven

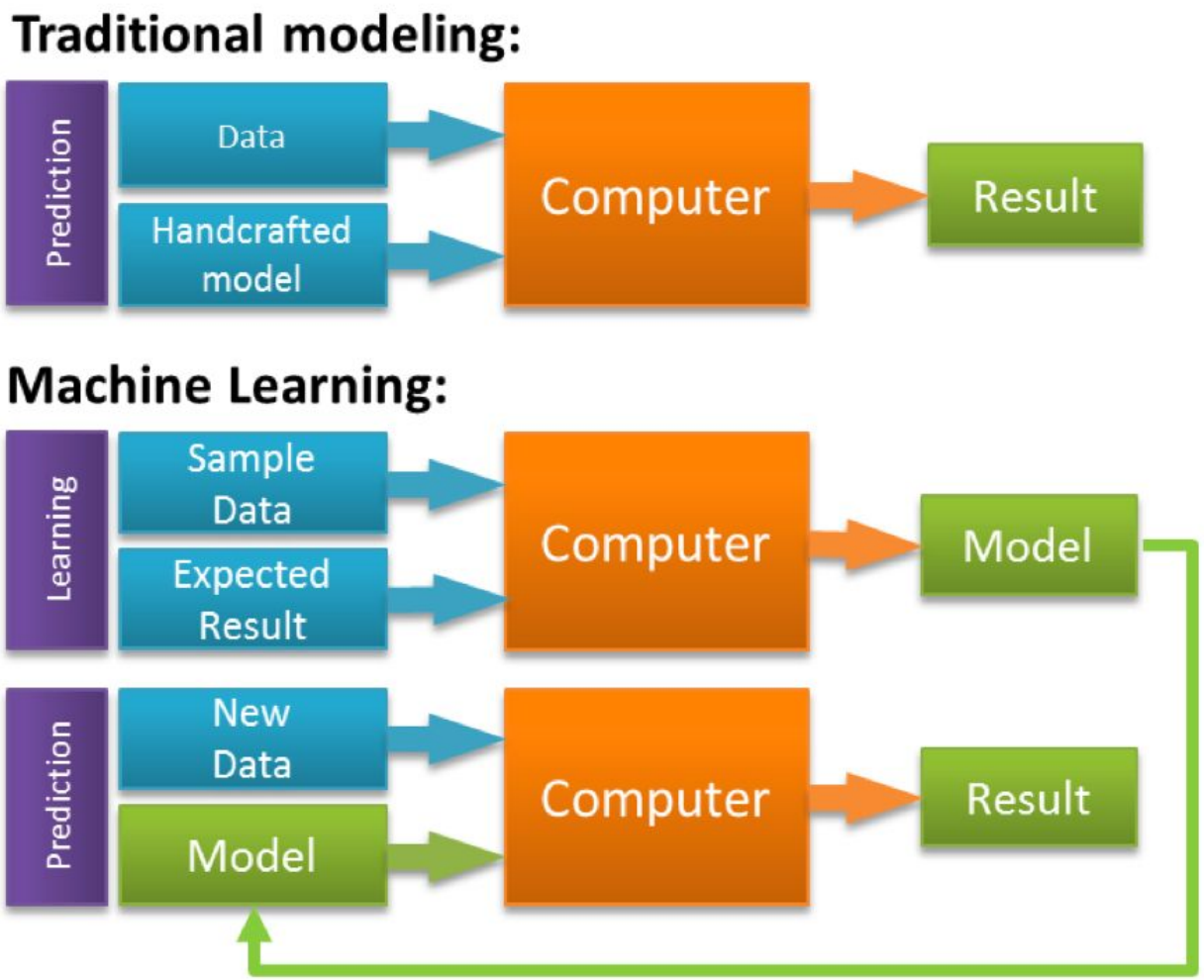
Course outline

- **Topics (Tentative):**
  - Day 01: Languages, Scripts, and Encoding of South Asian Languages.
  - Day 02: Phonology, Transliteration and Morphology of South Asian Languages.
  - Day 03: More on Morphology, Part of Speech and Multi-word tokenisation
  - Day 04: Syntax, Morphosyntax, and Semantics of South Asian Languages.
  - **Day 05: Machine/Deep Learning for South Asian Languages and winding up the course.**

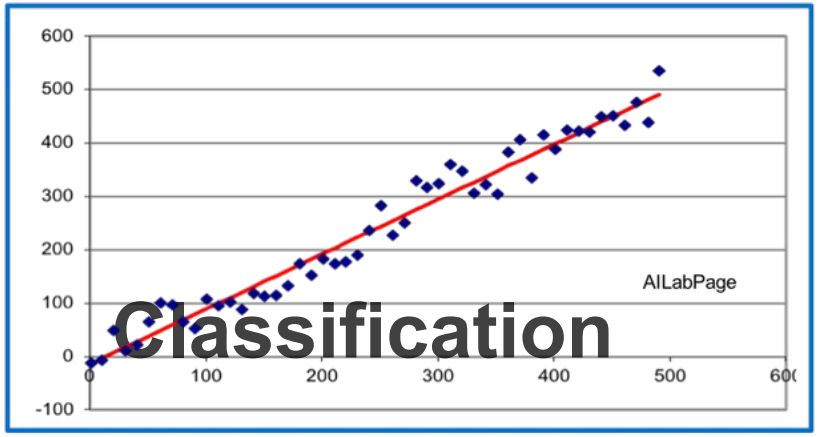
Machine/Deep Learning

Machine Learning

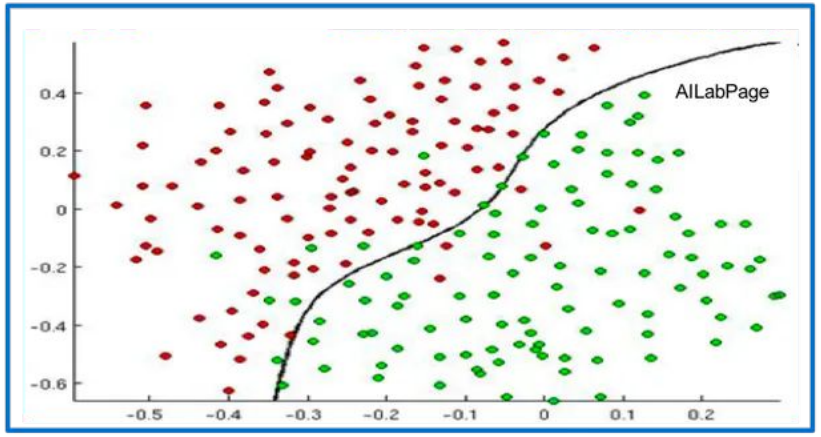




Supervised Learning



Classification



Regression

1. The system attempts to predict a value for an input based on past data.
2. Real number / Continuous numbers – Regression problem
3. Example – 1. Temperature for tomorrow

Classification

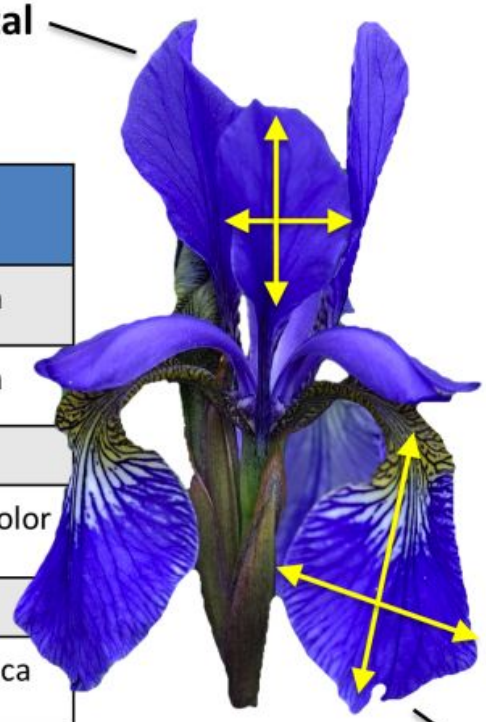
1. In classification, predictions are made by classifying them into different categories.
2. Discreate / categorical variable – Classification problem
3. Example – 1. Type of cancer 2. Cancer Y/N

Classification

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Petal



Sepal

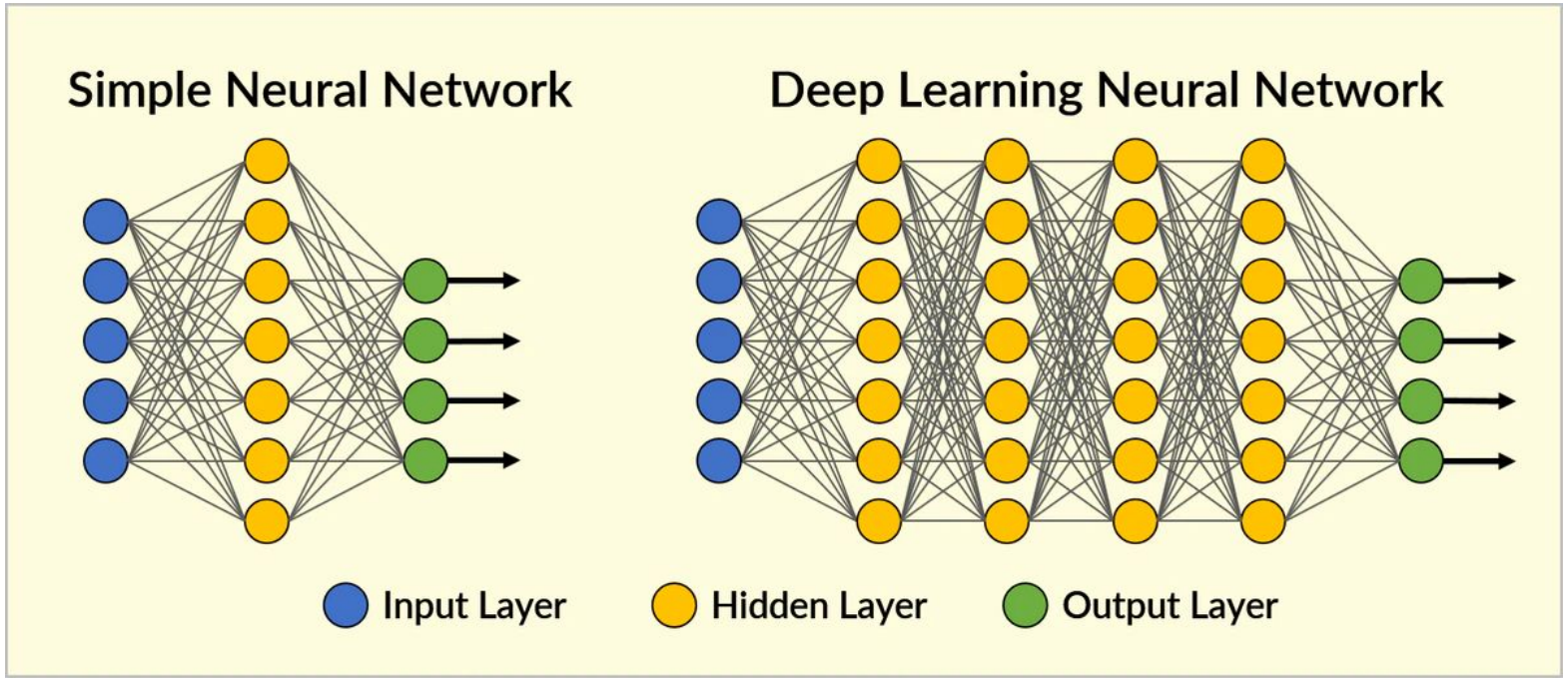
Class labels
(targets)

Features
(attributes, measurements, dimensions)

One challenge of apply machine learning on NLP problems is to represent the text by using numbers and vectors.

Classification

- **Neural Network**
  - Deep (Neural Network) Learning

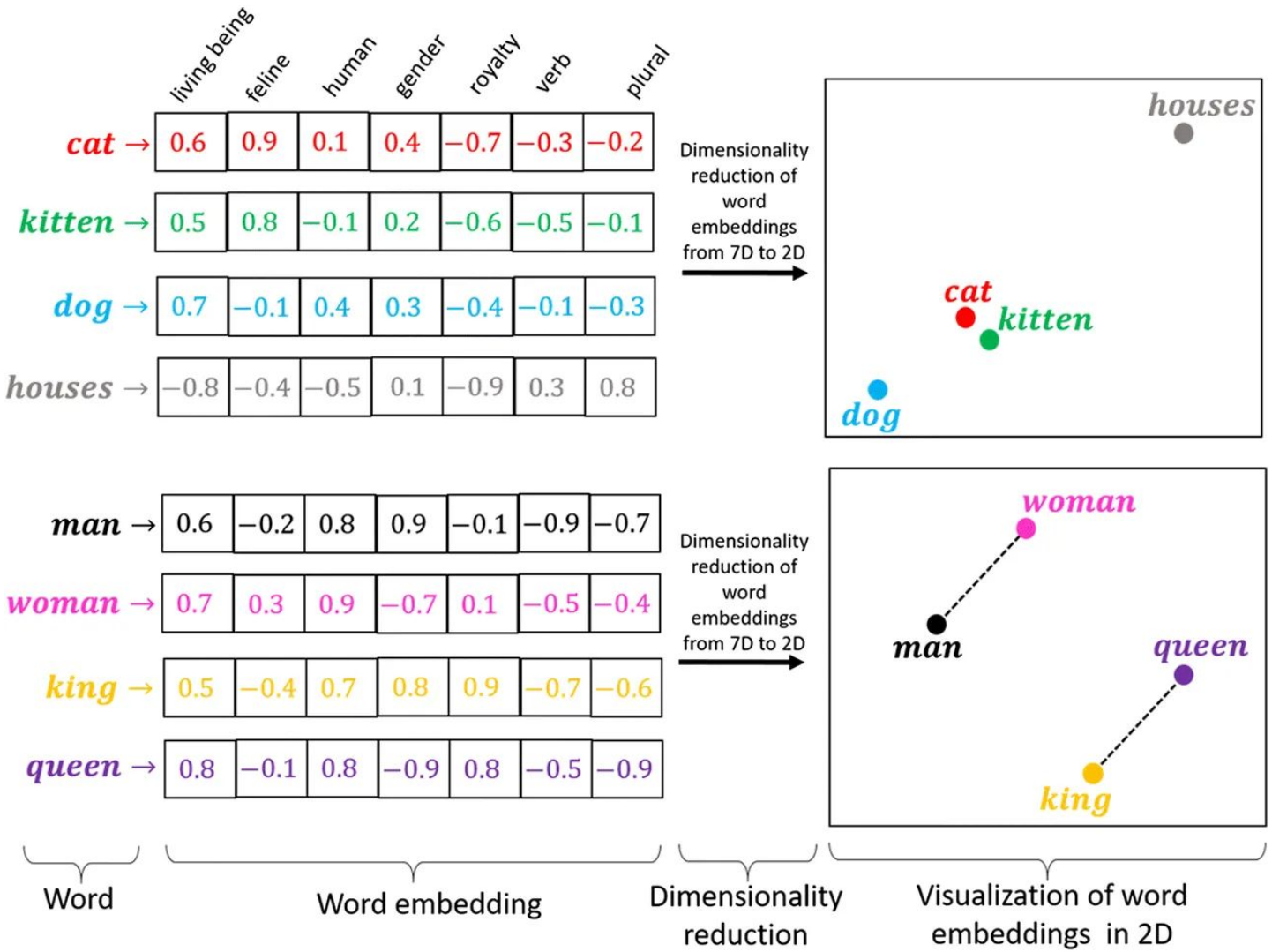


Feature Vector (ML using Bag of Word Approach)

Text Mining **deal**s *with* un**structur**ed **text**ual **inform**ation *and it* **discover**s *previously* un**known structur**e *and* **implicit mean**ings **buri**ed *within the* **large amount** *of* **text.** *A* **huge amount** *of* **inform**ation *is* **present** *as* un**structur**ed **text**, *so we* **need** *a* **special process** *to* **analyz**e *it.*

text	3+1
structur	2+1
inform	2
amount	2
min	1
deal	1
discover	1
previous	1
implicit	1
mean	1
bur	1
larg	1
hug	1
present	1
need	1
special	1
process	1
analyz	1

Word Embedding (Semantics oriented)



https://medium.com/@hari4om/word-embedding-d816f643140

Annotating the Corpora (for machine learning)

ذہین|Adj لڑکی|Noun نے|Adp روزانہ|Adv اچھی|Adj کتاب|Noun Verb|پڑھی

ہیں	بانی	کے	مائیکروسافٹ	گیٹس	بل	
Verb	Noun	AdP	Noun	Noun	Noun	POS
O	O	O	Org-B	Per-I	Per-B	IOB

dependency-conll - Notepad

File Edit Format View Help

1	ذہین	ذہین	Adj	Adj	-	2	amod
2	لڑکیاں	لڑکی	Noun	NN	-	6	subj
3	نے	نے	Adp	PP	-	2	case
4	اچھی	اچھا	Adj	Adj	-	5	amod
5	کتابیں	کتاب	Noun	NN	-	6	obj
6	پڑھیں	پڑھ	Verb	VB	-	0	ROOT
7	تھیں	ہے	Aux	Aux	-	6	aux

Challenges and Solutions

Low Resource Languages

- **Raw Text**
- **Annotating the Text**
  - resources in terms of time and money
  - finding/training skilled human resources
- **Finding/creating language specific annotation standards**
- **Creating/modifying the tools/architectures to deal with new annotation schemes**
- **Benchmarks**

Raw Text

Copyright and Sharing Issues is a hurdle.

- Common Crawl Corpus
  - petabytes of web data crawled since 2008.
  - Language identification of 160 languages
  - English 44%, Russian 6%, ……, Hindi 0.19% , Tamil 0.05%, Urdu 0.03% , Sindhi 0.002%
- Opus corpora
  - Parallel Corpora aligned manually or by text embeddings
- Corpora owned by organizations and societies

Annotated Text

Bag of Word Based Learning - Resources requiued

- **Normalization**
- **Tokenizers**
  - Multiword issues
  - Dealing (non-Latin) Punctuation marks
- **Stemmers/Lemmatizers**
- **Stop Words (with or without stemming)**

Types of (Transformer based) Deep Learning Solutions

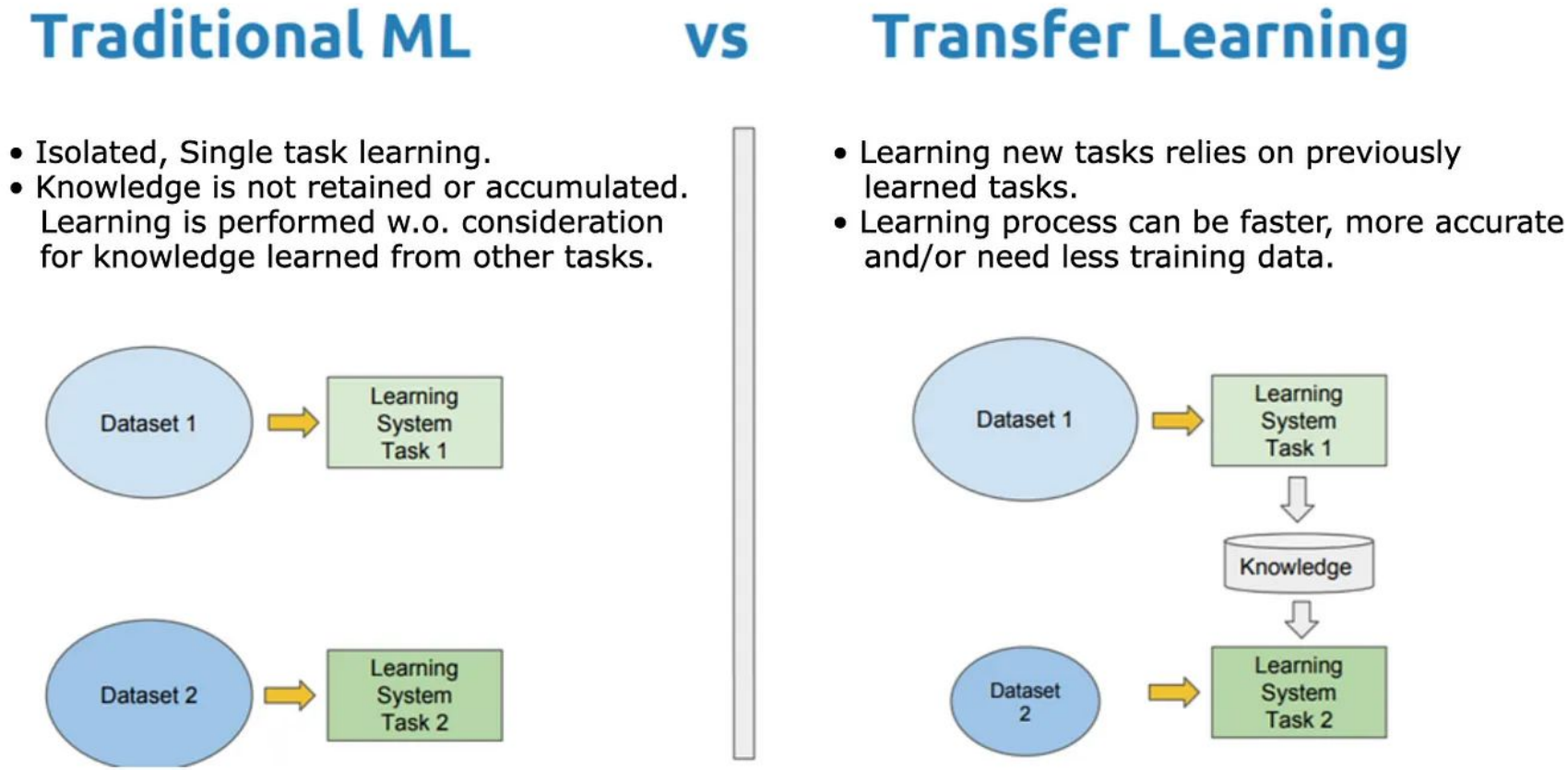
- **Zero shot Learning**
  Prompt:
  Give PoS (Part of Speech) tags for each word in the following {text}.
  {text} = I read a book

- **Few Shot Learning**
  Prompt:
  Give PoS (Part of Speech) tags for each word in the following {text}. Some examples of Pos Tagging are given below.
  {Input_1} = They bought few candies {Output_1} = They/PRON bought/VERB a/QUANT candies/NOUN
  {Input_2} = Children love candies {Output_2} = Children/PRON love/VERB candies/NOUN
  {text} = I read a book

Thousands or millions of examples of training data is not required. However, during the training the model should have seen many examples of that language.

Use Transfer Learning or Fine-tuning,if we have more annotated examples-

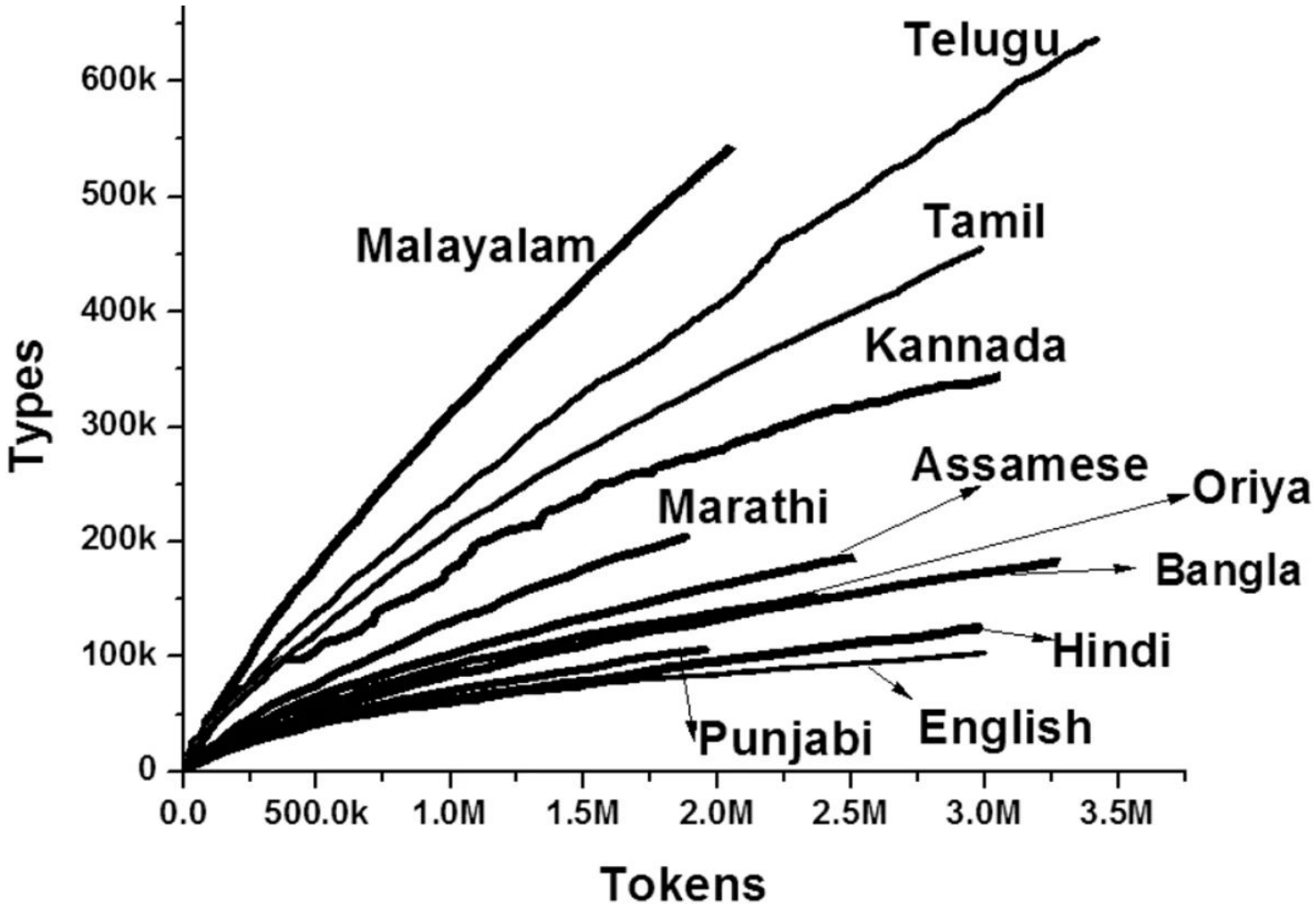
Transfer Learning



How Traditional ML differs from Transfer Learning

(Rich) Morphology

Rich Morpholgy



https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-023-00313-7/

Morphology Induction

- **Linguistica**
  - https://linguistica-uchicago.github.io/lxa5/
  - Unsupervised Learning from Corpus
  - Affixes, Signatures, and associated words
- **Morfessor**
  - https://morfessor.readthedocs.io/
  - Pre-trained segmentation models
  - Models can be trained

Examples are in the Python notebook

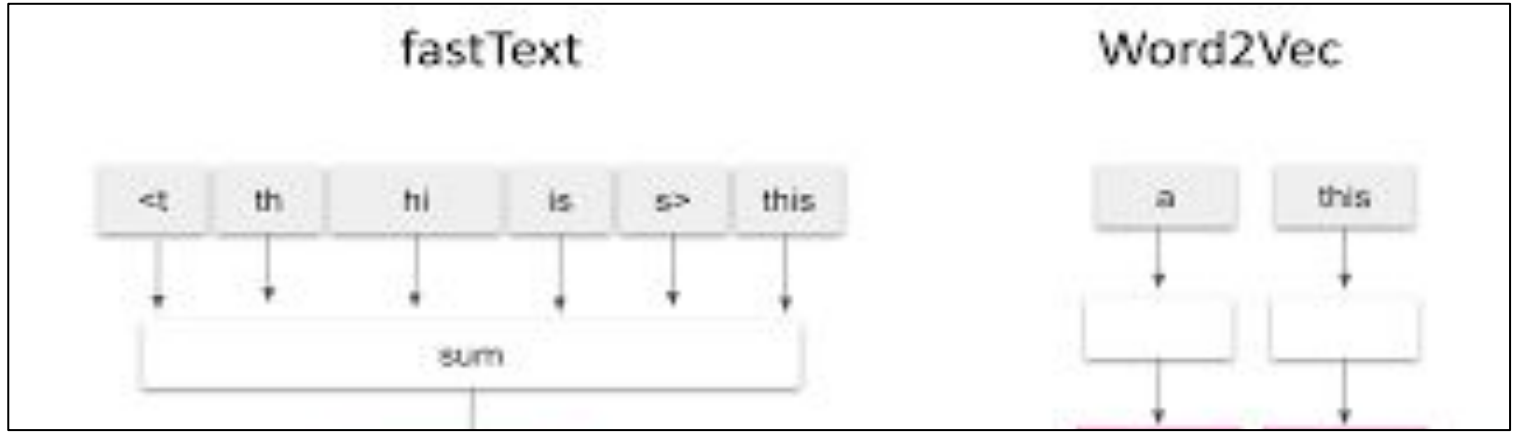
Morphology and Deep Learning

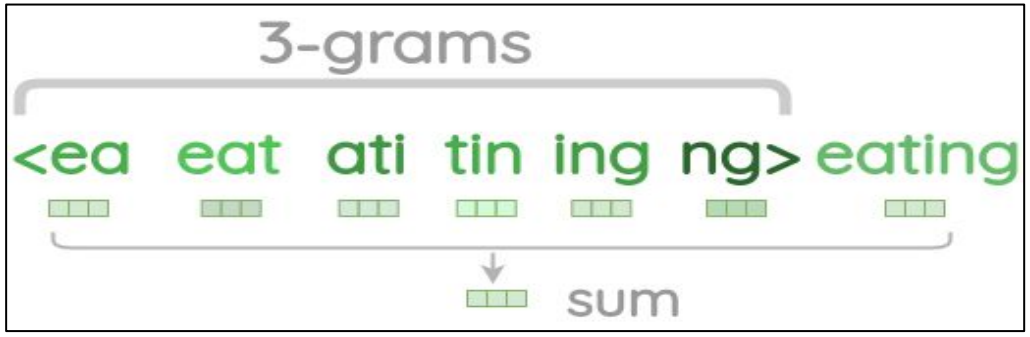
- Word Embedding
- Subword Tokenization

Word Embeddings

- **Word2vec**
  - learns embeddings of the words
  - Issue: morphological forms, spelling errors/variations

- **FastText**
  - Ngram based learning





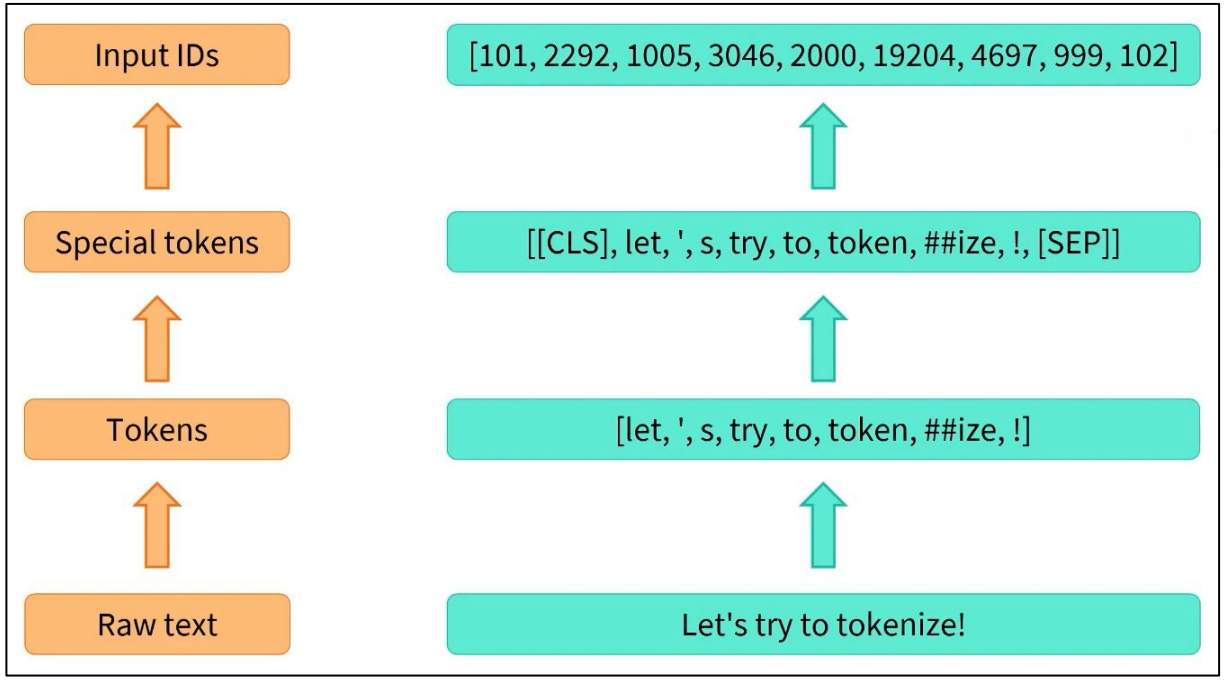
https://kavita-ganesan.com/fasttext-vs-word2vec/

https://amitness.com/2020/06/fasttext-embeddings/

Subword Tokenization

WordPiece

Byte Pair Encoding (BPE)



TikTokenizer

https://tiktokenizer.vercel.app/

Subword Tokenization: Issues

https://arxiv.org/pdf/2311.05845

Figure 1: Tokenizer comparisons between original LLaMA and Tamil LLaMA.

	Length	Content
Tamil Text	67	தமிழ், உலகில் உள்ள முதன்மையான மொழிகளில் ஒன்றும் செம்மொழியும் ஆகும்.
LLaMA-2 Tokenizer	89	'<s>', '', 'த', 'ம', 'ி', 0xE0, 0xAE, 0xB4, 'ழ்', ',', '', 0xE0, 0xAE, 0x89, 'ல', 'க', 'ி', 'ல', 'ள்', '', 0xE0, 0xAE, 0x89, 'ள', 'ள்', 'ள', '', 'ம', 'ா', 'த', 'ன', 'ம்', 'ம', 'ை', 'ய', 'ா', 'ன', '', 'ம', 0xE0, 0xAF, 0x8A, 0xE0, 0xAE, 0xB4, 'ி', 'க', 'ள', 'ி', 'ல', 'ள்', '', 0xE0, 0xAE, 0x92, 'ன', 'ற்', 'ற', 'ம', 'ம்', '', 'ச', 0xE0, 0xAF, 0x86, 'ம', 'ம்', 'ம',0xE0, 0xAF, 0x8A, 0xE0, 0xAE, 0xB4, 'ி', 'ய', 'ா', 'ம', 'ம்', '', 0xE0, 0xAE, 0x86, 'க', 'ா', 'ம', 'ம்', '.'
Tamil LLaMA Tokenizer	18	'<s>', 'தம', 'ி', 'ழ்', ',', 'உ', 'ல', 'கில்','உள்ள', 'முதன்மையான', 'மொழிகளில்', 'ஒன்றும்', 'செம', 'ம்', 'மொழி', 'யும்', 'ஆகும்', '.'

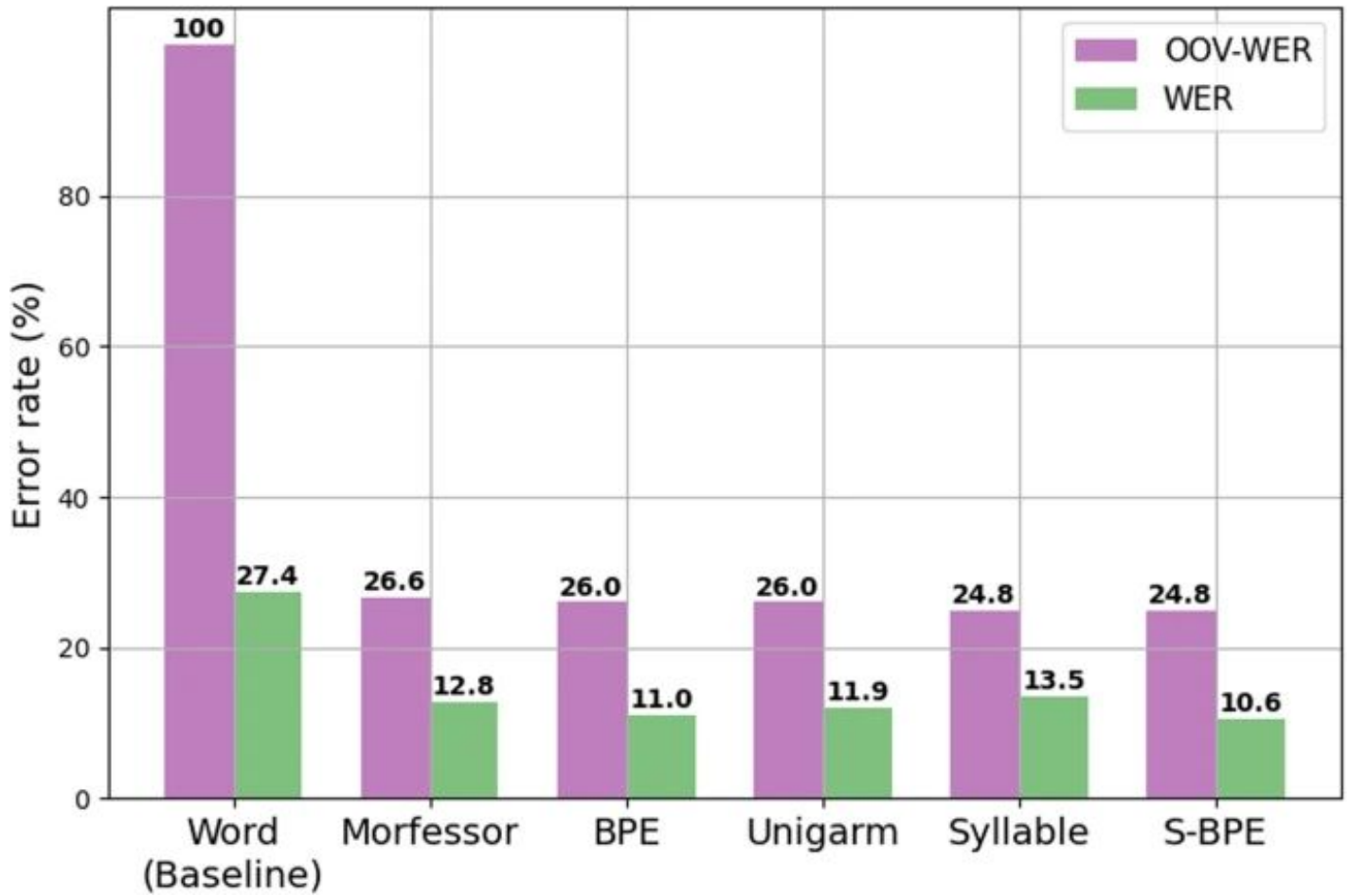
Morph Aware Subword Tokenization

- Creating morphology based tokens before subword-learning can give better results.
- A similar work is:

  Improving speech recognition systems for the morphologically complex Malayalam language using subword tokens for language modeling

  https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-023-00313-7

Syllable Aware Subword Tokenization



Thank you

Kengatharaiyer Sarveswaran
University of Jaffna, Sri Lanka.
University of Konstanz, Germany
sarves@univ.jfn.ac.lk
sarves.github.io

Tafseer Ahmed
Senior NLP Scientist, Alexa Translations
tafseer@gmail.com