bangla2vec

A Language Model for Bengali

Let’s Talk About Languages

- If you are multilingual, you may have noticed that the different languages you speak will make you stretch in different ways
- Different languages can also shape the way you think and perceive the world around you
- Languages can bring about innovations

Natural Language is Complex

- How words acquire meaning?
    - Why ***C-A-T*** refers to cat, an animal?
    - What makes ***C-A-T*** in that order makes it refer to
        - cat, that meows,
        - and NOT a dog, that barks.
    - Are the meaning of words entirely composed from the individual letters they contain?
- How can meaning emerge from a series of words?
    - Why (1) makes sense where (2) doesn’t,
        1. *How are you?*
        2. *Are how you?*
- No one understands, how we understand language.
    - Even though the following statement is syntactically correct, why does it read absurd?
        - *the green marble went to sleep last year.*

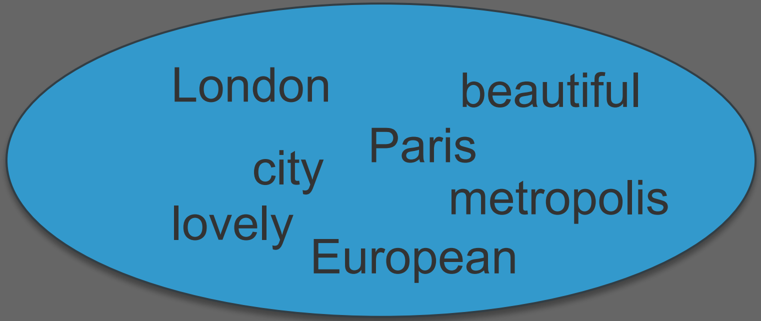
How can we actually *Teach* a Computer Language?

- Computers are really good at crunching numbers, but not so much with text. We have to specifically instruct them on how to handle text.
- For text analytics to be carried out, we have to represent the text in a form that the computers can understand, i.e., in the form of numbers
- After this is done, analysis can be carried out on these numbers in the same way as dealing with numbers

Bag of Words Model

“London is a beautiful European city. Paris is a lovely European metropolis.”

- For the above text, the extracted words are:
- *“London”, “is”, “a”, “beautiful”, “city”, “Paris”, “lovely”, “metropolis”, “European”*
- Usually, words like “is”, “a”, etc. are ignored, since they are only used for grammar.
- So, we finally obtain a “Bag of Words” –
- These are the words that will be used.



London

beautiful

Paris

city

metropolis

lovely

European

Disadvantage of this model

- It says nothing about the order of the words in the original text.
- It says nothing about the context of the text.
- It says nothing about the meanings of the words.

For example, the computer thinks the words “*city*”, “*lovely*”, “*metropolis*” are equivalent, although “*city*” and “*metropolis*” are mostly equivalent, and “*lovely*” is something very different.

London	beautiful	city	Paris	lovely	metropolis	European
1	1	1	1	1	1	2

Word2Vec

“You shall know a word by the company it keeps” - J.R. Firth

What does *word2vec* do?

- Unsupervised learning approach (Auto Encoder)
- It looks at the text, and assigns each word a vector of a fixed size.

How does it do this?

- When given the text, it looks at each word and the words around it.
- In this way, it trains itself on the text, and recognizes the order of each word, and the structure of the sentences.
- At the end of training, each word is represented by an N-dimensional vector, where N is typically in the hundreds.

“London is a beautiful European city. Paris is a lovely European metropolis.”

Guess the Meaning of the Word

The cat is *zigmoiding* on the table

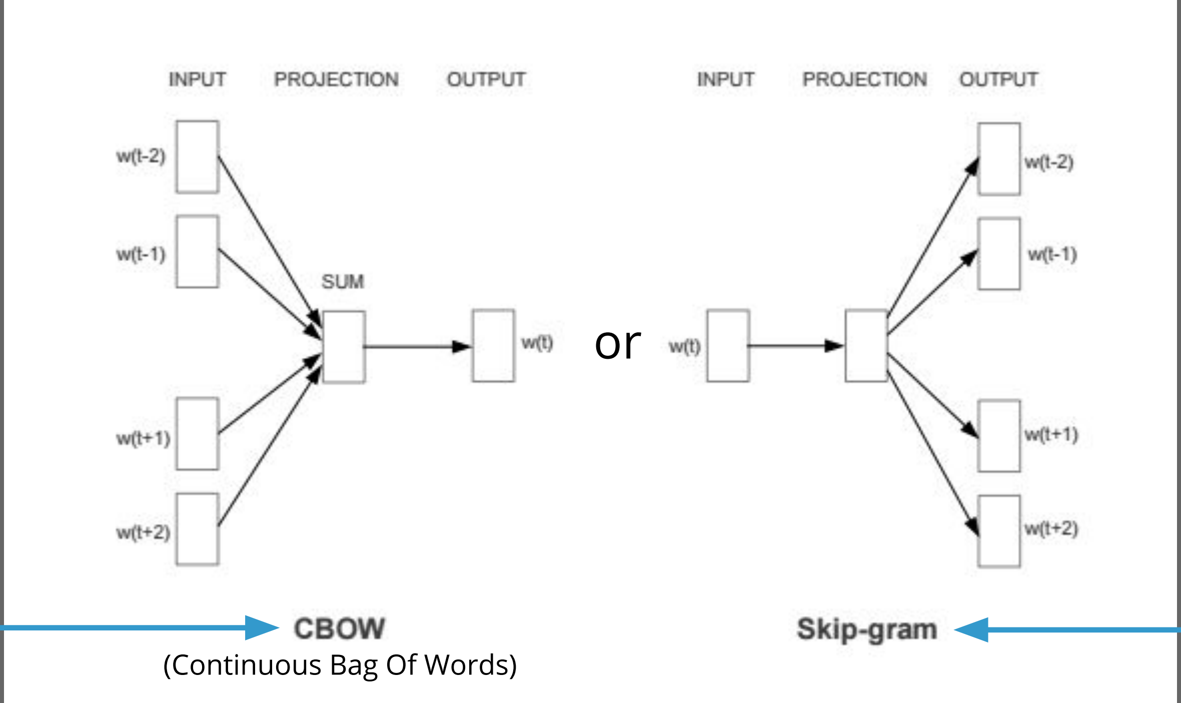
Someone is *zigmoiding* in my place in the office

Welcome! Please *zigmoid* down.

What does *zigmoid* mean?

word2vec – Word to Vector

Each words is analyzed in the context of the words around it. There are two ways of doing this:



or

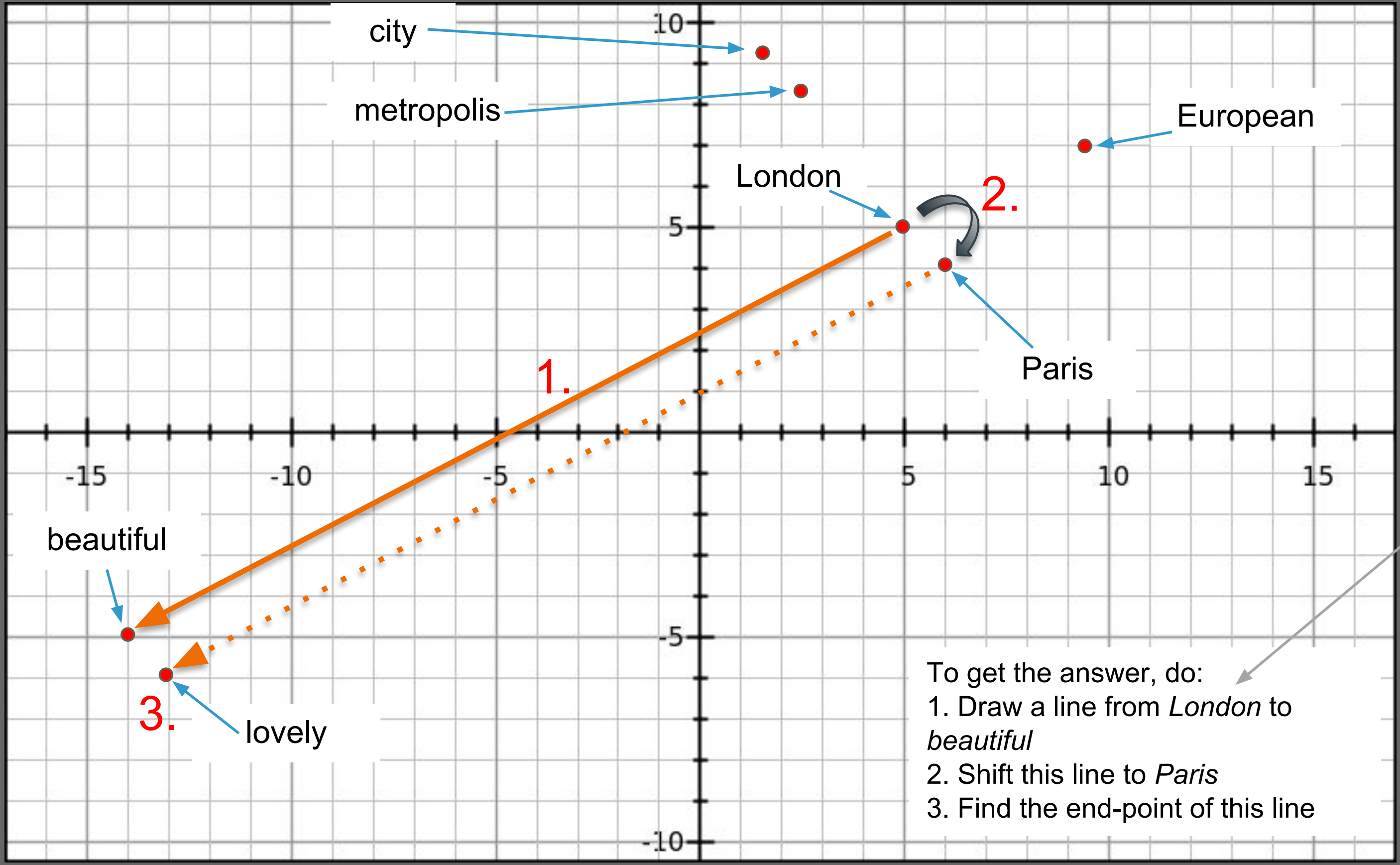
CBOW
(Continuous Bag Of Words)

Skip-gram

Given a set of (neighboring) words, **guess single words** that potentially occur along with this set of words.

Guess potential neighboring words based on the single word being analyzed.

word2vec – Word to Vector



To get the answer, do:
1. Draw a line from *London* to *beautiful*
2. Shift this line to *Paris*
3. Find the end-point of this line

From the relative positions, you can observe analogous relationships too!

The computer knows the answer to:
If
London : *beautiful*
then
Paris : *????????*

Answer
lovely

Similarly,
If
London : *city*
then
Paris : *metropolis*

Bangla2Vec

- Using Word2Vec on Bengali
- Data
  - Crawling Bengali News Websites to get Tagged News Data
  - Bengali Wikipedia Dump
  - Bengali CommonCrawl - In the works
- Training
  - Gensim Word2Vec
  - News Classifier using pre-trained embeddings
- Demo - http://indicnlp.meain.io

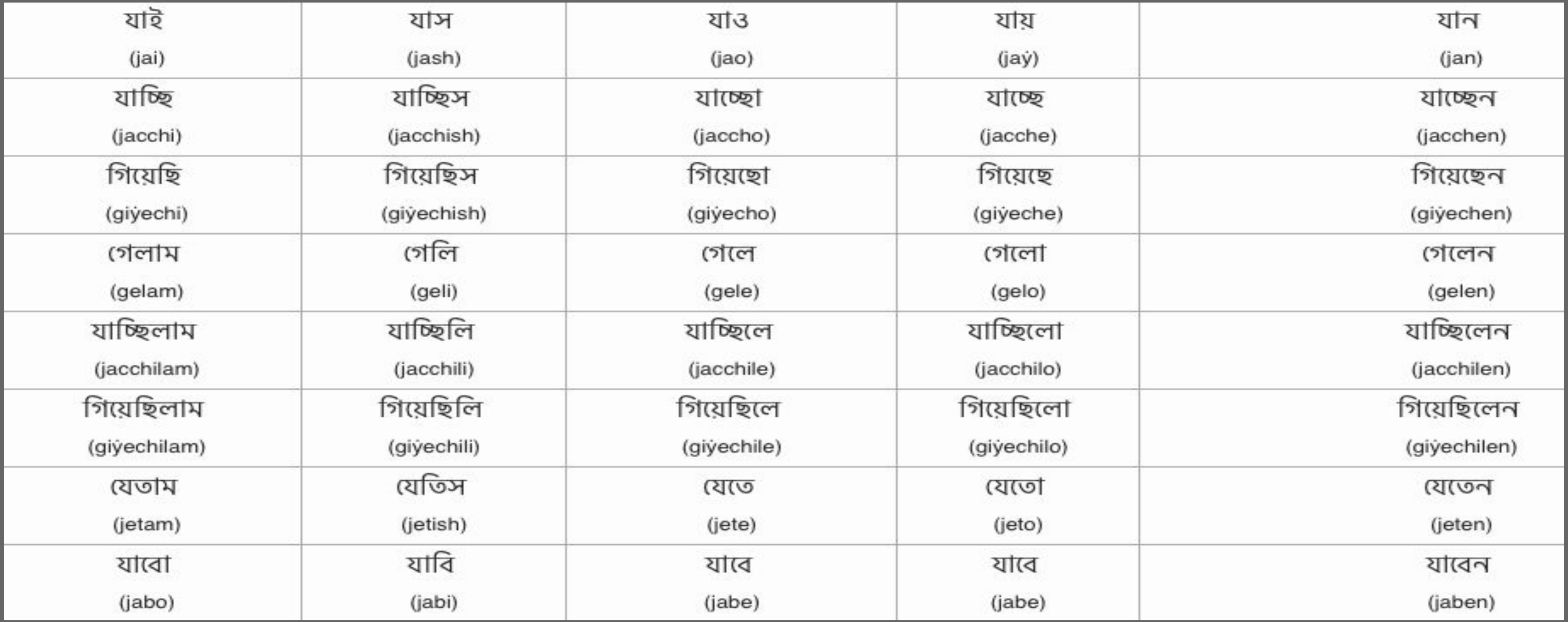
Existing Work

- Embeddings
  - FastText Bengali Embeddings: Trained on Wikipedia dump and CommonCrawl
  - Google Translate (Not open Source)
- Data
  - OPUS: Open Parallel Corpus
  - LDCIL, CIIL, IILS (Not open source)
- Research/Researchers
  - Sudeshna Sarkar: Hidden Markov Models, Stemming
  - Utpal Garain: Lemmatizer, POS Tagger, Translation
  - BanglaNLP Group
- BanglaKit (github)
- Bengali.ai

Challenges

- Data Pre-processing
  - How to deal with tenses, half-letters, inflections and references?
  - How to deal with a word that has two meanings? : ULMFIT
- Training: Getting proper Word Embeddings
- Data Collection
  - Different dialects and scripts
  - Getting tagged data: For classification, NER, Translation
  - Data should be representative
- Data Curation
  - What is the best way to store and distribute our data?
- Getting Different Types of Data
  - Voice/Sound data
  - Fish or Sweet Classifier?

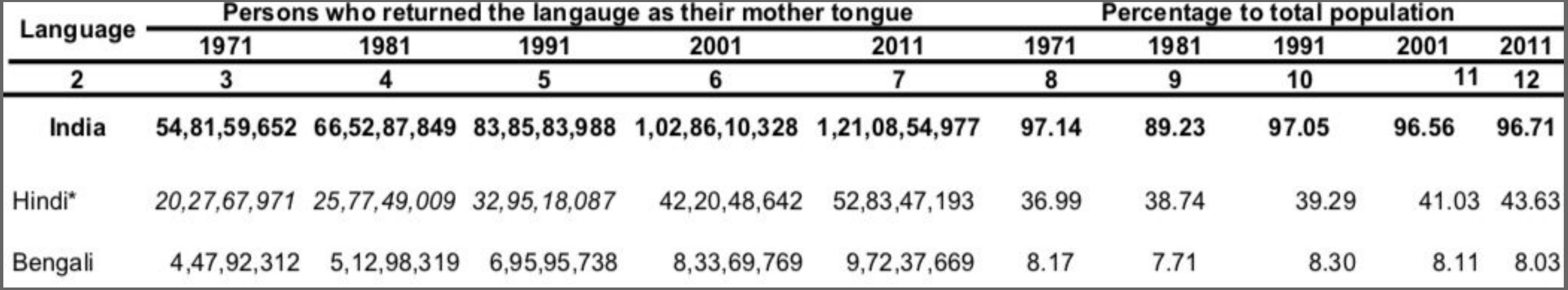
40 Different Forms of “যাওয়া”



যাই (jai)	যাস (jash)	যাও (jao)	যায় (jaẏ)	যান (jan)
যাচ্ছি (jacchi)	যাচ্ছিস (jacchish)	যাচ্ছো (jaccho)	যাচ্ছে (jacche)	যাচ্ছেন (jacchen)
গিয়েছি (giẏechi)	গিয়েছিস (giẏechish)	গিয়েছো (giẏecho)	গিয়েছে (giẏeche)	গিয়েছেন (giẏechen)
গেলাম (gelam)	গেলি (geli)	গেলে (gele)	গেলো (gelo)	গেলেন (gelen)
যাচ্ছিলাম (jacchilam)	যাচ্ছিলি (jacchili)	যাচ্ছিলে (jacchile)	যাচ্ছিলো (jacchilo)	যাচ্ছিলেন (jacchilen)
গিয়েছিলাম (giẏechilam)	গিয়েছিলি (giẏechili)	গিয়েছিলে (giẏechile)	গিয়েছিলো (giẏechilo)	গিয়েছিলেন (giẏechilen)
যেতাম (jetam)	যেতিস (jetish)	যেতে (jete)	যেতো (jeto)	যেতেন (jeten)
যাবো (jabo)	যাবি (jabi)	যাবে (jabe)	যাবে (jabe)	যাবেন (jaben)

Why this Work is Important

- Open Source: There is a lot data available, but none are open source
- Bengali is a ~~dying~~ changing language (Indian Language Census 2011)
- We need to preserve less spoken dialects and scripts as they are all unique and contain a wealth of information
- Bangla2Vec and IndicNLP can help do this



Language	Persons who returned the langauge as their mother tongue					Percentage to total population				
	1971	1981	1991	2001	2011	1971	1981	1991	2001	2011
2	3	4	5	6	7	8	9	10	11	12
India	**54,81,59,652**	**66,52,87,849**	**83,85,83,988**	**1,02,86,10,328**	**1,21,08,54,977**	**97.14**	**89.23**	**97.05**	**96.56**	**96.71**
Hindi*	*20,27,67,971*	*25,77,49,009*	*32,95,18,087*	42,20,48,642	52,83,47,193	36.99	38.74	39.29	41.03	43.63
Bengali	4,47,92,312	5,12,98,319	6,95,95,738	8,33,69,769	9,72,37,669	8.17	7.71	8.30	8.11	8.03

How You Can Contribute

- If you're a linguist, you can help design news tasks for your mother tongue.
- If you're a machine learning enthusiast, you can build and test new models for the tasks of your interest
- If you are a programmer, you can help build intuitive interfaces for tagging
  - Web App
  - Android/iOS App
- Python Package for IndicNLP
- If you know English and Indic languages, you can help us translate.
- If you know multiple Indic languages, you can help translate across indian languages
- Most importantly, you can help us Acquire, Tag and Curate datasets!

IndicNLP

- In active development
  - Malayalam: Adam and Kamal; Winner ICFOSS Kerala
  - Tamil: Selva
  - Bengali: Anirban, Soham
  - Hindi: Archana
  - Python Package: pip install indicnlp
  - Docs and Website: Abin, Pal, Selva
- In the works
  - Gujarati
  - Punjabi
  - Sanskrit

www.indicnlp.github.io

Contact Us

- **Feedback- https://tinyurl.com/indicnlptalk**
- **Contact Me:**
- 96soham96@gmail.com
- csoham.com
- Project Details:
  - Indicnlp.github.io
  - Indicnlp.meain.io
  - https://groups.google.com/forum/#!forum/indicnlp