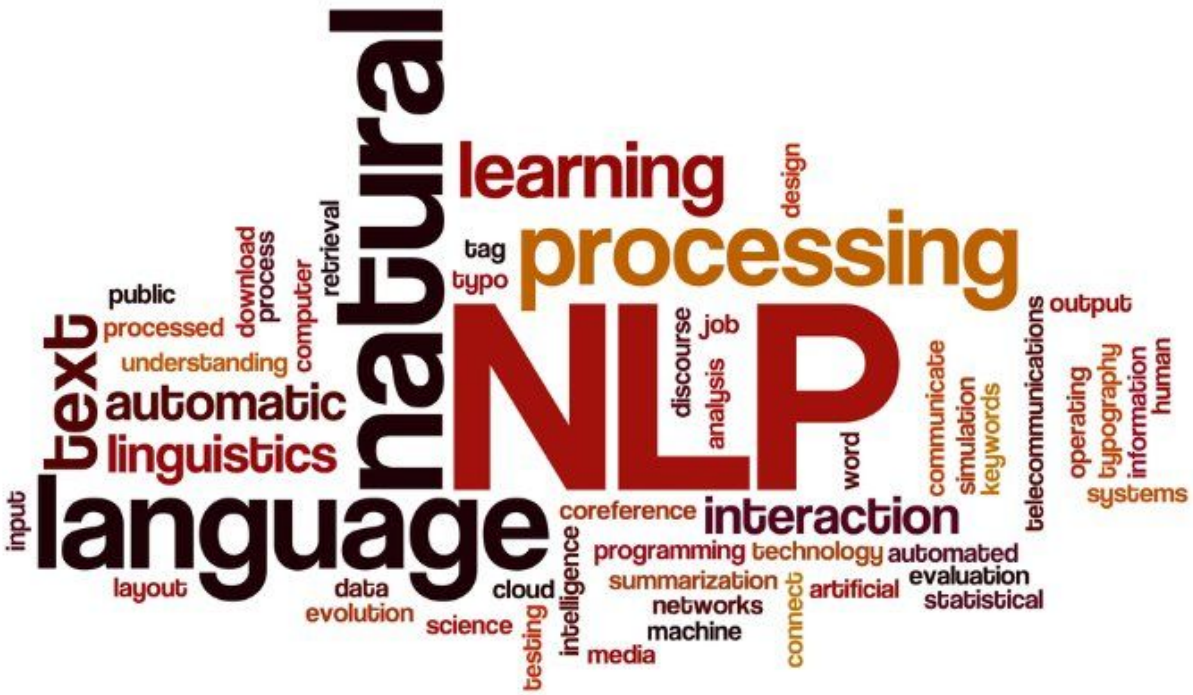
NATURAL LANGUAGE PROCESSING
CODE MIXED NATURAL LANGUAGE GENERATION

Dinesh Kakarla, Vamshi Krishna srirangam



Introduction

Now a days information in social networks, blogs, microblogs is a major form of raw data for extracting information and developing language technologies. But the problem with social networks is that people tend to chat in different languages. Like in India we mix languages while we chat, India being an abode for many languages people tend to speak a language by inserting english words into the sentences at times, at times phonetic typing/ roman script/ transliteration, anglicisms also. While english is still being the principal language for communication there is a growing need to develop

technologies for other widely spoken languages too.. for example in India we can consider Hindi, bengali, telugu etc.. There are many applications to be developed on this like language recognition, POS tagging, text processing etc.

Challenge

Our aim in this project is to generate sentences which have mixed languages.

This information for sake of various applications.

What we have:

We are given data which contains sentences that contain words of different languages like

English-hindi

English-telugu

English-bengali

These mixing of languages is prominent in countries like india.

We have collected this data from facebook, whatsapp, twitter.

What we need:

As of this project we need to generate the sentences which are mixture of more than one language.

What we have to do:

- Extract information from data.
- Convert input into useful format
- Do language modelling on the words and on their associated tags, basically n-gram modelling
- Use this information from n-gram modelling to go to further developments
- Use probabilities measures for the sake of generation of sentences.
- Select the unigram which has maximum occurrences, then search for all the bigrams with the selected unigram at the beginning, suppose the selected unigram is “the”, now collect all bigrams starting with “the” for e.g they are
  The boy
  The car
  The girl etc..
  Then select the one that has max occurrences or probability, say it “the boy” now repeat the same procedure for trigrams and so on..
  Ideally up to 6 grams.

- This is the way to generate sentences using words
- You can also generate sentences using tag sequences
- Consider a tag sequence containing three tags as a fragment
- Generate fragments from sentences
- Calculate all the possible combinations of fragments, i.e suppose (A, B, C) is a fragment..

  Consider all the possible combinations and calculate their occurrences i.e

  (A, B, C)(X, Y, Z)

  (A, B, C)(P, Q, R)

  (A, B, C)(U, V, W)

  Select the one with max occurring say it is (A, B, C)(X, Y, Z)
- Now calculate the tag given word probabilities i.e for a given tag their exist's many word sequences in the corpus for e.g

  (DT, NOUN, VERB) there exists

  The boy came

  The girl gone

  The camera clicked etc..

  Select the one that has many occurrences, for example "the boy came". Now the selected sentence is "the boy came". In this way generate the sentences
- Apply smoothing so that the same word is not selected every time, i.e apply some function(e^x) or multiply with 0.9 etc..
- This way you generate the required sentences.

********** **********