SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 97
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Artificial Intelligence and Data Science

Course Name – 19AD601 – Natural Language Processing

III Year / VI Semester

Unit 1 – Introduction

Topic 2- Language Model

Language Model

Language modeling (LM) is the use of various statistical and probabilistic techniques to determine the probability of a given sequence of words occurring in a sentence. Language models analyze bodies of text data to provide a basis for their word predictions.

It is widely used in predictive text input systems, speech recognition, machine translation, spelling correction etc. The input to a language model is usually a training set of example sentences. The output is a probability distribution over sequences of words.

Types of Language Model

- Grammar-based models
- Statistical models

Language Model

Grammar based Model

Grammar contains Symbols, Rules, Procedure of rule application.

Formal grammar

More technically, a formal grammar consists of a finite set of terminal symbols, a finite set of nonterminal symbols, a set of rules (also called production rules) with a left- and a right-handed side, each consisting of a word a start symbol.

Formal grammars usually have two special symbols

- S: the start symbol
- ε: the empty string (sometimes: λ)

Language Model

Formal definition

A grammar G = ⟨Φ, ∑, R, S⟩ consists of,

- An alphabet Φ of nonterminal symbols,
- An alphabet ∑ of terminal symbols,
- A set R ⊆ Γ*╳Γ* of rules (where Γ = Φ ∪ ∑),
- A start symbol S ∈ Φ

Representing formal grammar

- Nonterminals are usually represented by upper-case letters {S, A, B}
- Terminals by lower case letters {a, b, c}
- The start symbols by S

Language Model

Chomsky hierarchy

- 4 types of grammars (Type-0 to Type-3)
- Type-0: recursively enumerable
- Type-1: context sensitive
- Type-2: context free (CFG)
- Type-3: regular

Type-0: recursively enumerable

- All grammars and languages
- Those that can be recognised by a Turing Machine
- Pattern:

$\alpha \rightarrow \beta$

(where α and β are any string of terminals and nonterminals, including the empty string)

Language Model

Statistical Language Models

Statistical models include the development of probabilistic models that are able to predict the next word in the sequence, given the words that precede it.

1.N-Gram: This is one of the simplest approaches to language modelling.

Here, a probability distribution for a sequence of ‘n’ is created, where ‘n’ can be any number and defines the size of the gram (or sequence of words being assigned a probability

2.Unigram: The unigram is the simplest type of language model. It doesn't look at any conditioning context in its calculations. It evaluates each word or term independently.)

Language Model

3.Bidirectional: Unlike n-gram models, which analyze text in one direction (backwards), bidirectional models analyze text in both directions, backwards and forwards.

4.Exponential: This type of statistical model evaluates text by using an equation which is a combination of n-grams and feature functions.

5.Continuous Space: In this type of statistical model, words are arranged as a non-linear combination of weights in a neural network.

Language Model

Applications of statistical language modeling

1.Statistical language models are used to generate text in many similar natural language processing tasks, such as:

2.Speech Recognition - Voice assistants such as Siri and Alexa are examples of how language models help machines in processing speech audio.

3.Machine Translation - Google Translator and Microsoft Translate are examples of how NLP models can help in translating one language to another.

4.Sentiment Analysis - This helps in analyzing sentiments behind a phrase.

5.Text Suggestions - Google services such as Gmail or Google Docs use language models to help users get text suggestions while they compose an email or create long text documents, respectively.

6.Parsing Tools - Parsing involves analyzing sentences or words that comply with syntax or grammar rules. Spell checking tools are perfect examples of language modelling and parsing.

THANK YOU