SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 007
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 – 16AD601 – Natural Language Processing

III Year / VI Semester

Unit 1 – Introduction

Topic 6- Detecting and Correcting Spelling Errors

Detecting and Correcting Spelling Errors

Spelling Correction is a very important task in Natural Language Processing. It is used in various tasks like search engines, sentiment analysis, text summarization, etc.

As the name suggests, we try to detect and correct spelling errors in spelling correction. In real-world NLP tasks, we often deal with data having typos, and their spelling correction comes to the rescue to improve model performance.

Spelling checking in used in various applications like machine translation, search, information retrieval etc. Spell checking technique comprises of two stages

i. Error detection and

ii. Error correction

Detecting and Correcting Spelling Errors

TYPES OF SPELL ERRORS

Various techniques that were designed on the basis of spelling errors and trends also called error patterns, and most notable studies on these were performed by Damerau. According to these studies spelling errors are generally divided into two types

- Typographic errors and
- Cognitive errors.

Typographic errors (Non Word Errors): These errors occur when the correct spelling of the word is known but the word is mistyped by mistake

Cognitive errors (Real Word Errors): These errors occur when the correct spellings of the word are not known. In the case of cognitive errors, the pronunciation of misspelled word is the same or similar to the pronunciation of the intended correct word

Detecting and Correcting Spelling Errors

ERROR DETECTION

For error detection each word in a sentence or paragraph is tokenized by using a tokenizer and checked for its validity.

The candidate word is a valid if it has a meaning else it is a non word. Two commonly used techniques for error detection is

- N-gram analysis and
- Dictionary/Wordnet lookup.

Detecting and Correcting Spelling Errors

N-gram Analysis

N-gram analysis is a method to find incorrectly spelled words in a mass of text. Instead of comparing each entire word in a text to a dictionary, just ngrams are checked.

A check is done by using an n-dimensional matrix where real n-gram frequencies are stored. If a non-existent or rare n-gram is found the word is flagged as a misspelling, otherwise not.

An n-gram is a set of consecutive characters taken from a string with a length of whatever n is set to.

Dictionary Lookup

A dictionary/Wordnet is a lexical source that contains list of correct words a particular language.

The non-word errors can be easily detected by checking each word against a dictionary.

Detecting and Correcting Spelling Errors

ERROR CORRECTION

Error correction consists of two steps:

The generation of candidate corrections: The candidate generation process usually makes use of a precompiled table of legal n-grams to locate one or more potential correction terms.

The ranking of candidate corrections: The ranking process usually invokes some lexical similarity measure between the misspelled string and the candidates or a probabilistic estimate of the likelihood of the correction to rank order the candidates.

THANK YOU