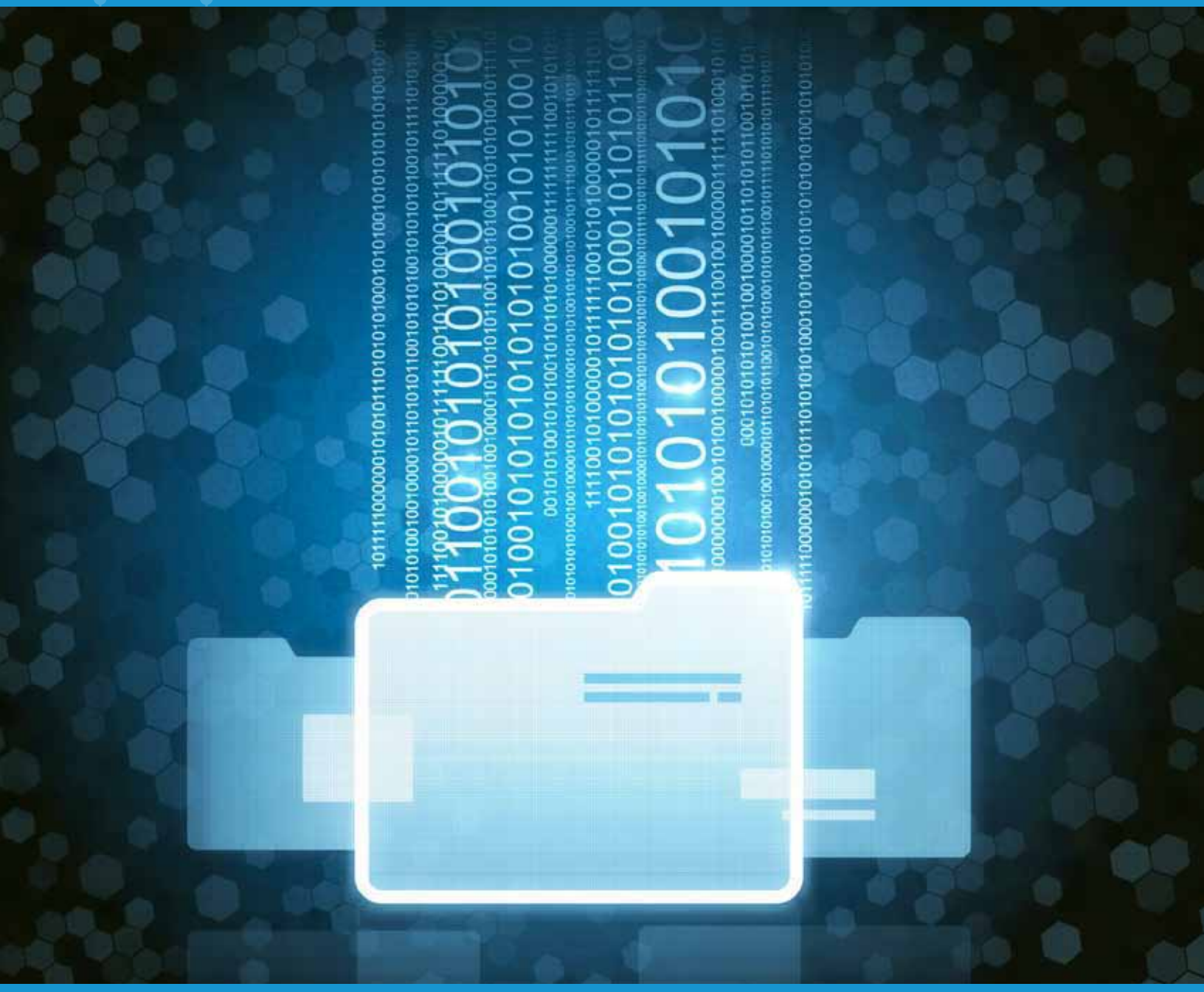
Mphasis
The Next Applied



Deep learning for NLP based Context Specific Spelling Error Correction

Ashutosh Vyas
Assistant Manager, Mphasis NEXTlabs

NEXT labs

Introduction

In today's world, lot of information is shared via digital documentation, which can be financial reports, annual reports, insurance policy documents, medical reports etc. Majority of these documents are manually created by professionals, because of which there can be spelling mistakes induced while building the documents. This misspelled words are a great hindrance in further processing and understanding of the documents. This white paper explores the different kind of approaches to this problem in an efficient way.

Abstract

This paper presents how deep learning can be used for *context based spelling error correction*, and highlights the existing methodologies and approaches available to fix the spelling errors.

Keywords: NLP: natural language processing, NN: Neural networks, ANN: artificial neural networks, RNN: recurrent neural networks,

MM: Markov models. LSTM: long-short-term-memory.

Why spelling error a critical problem?

Spelling mistakes caused due to human intervention is a critical problem because of following reasons:

a) Create problems in understanding the correct meaning of the text

b) Can be responsible for conveying different meaning of the text altogether

c) Hinder process automation as majority of text and document processing match words, and misspelled words can create problems in smooth functioning of data extraction and information retrieval algorithms

The above-mentioned problems make spelling error correction a major task that should be resolved before further processing of the digital data.

Concept of Solution

Spelling errors could be caused due to misspelled words with:

a) One or two missing letters

b) Wrong letter used, for e.g. *hello as gello*

c) Jumbled letters for example: *Letters as eltters*

Each misspelled word can be replaced with multiple words that are available in the dictionary.

Our task is to identify the exact word which best suits the given context.

This can be done in two steps:

1) Creating all possible correct dictionary words that can replace the misspelled word. This is known as the Cartesian product.

2) Identifying the best suitable word from the above set of words

Creating the Cartesian

Let us now look at the methods to develop a valid set of all the possible replacement words that can exist, known as the Cartesian set.

Replacing single character

Replace a single character from the misspelled word by all the letters from A-Z and try to match the newly created words with the correct word from dictionary.

Adding 2 character

Add one character each in the beginning and at the end to make all possible combinations of correct dictionary words. These 2 characters are formed by using letters from A-Z.

Note: Here we are adding two characters based on the assumption that this can be the maximum number of characters a human may forget while typing the correct word.

Cosine Similarity

Cosine similarity is the distance between two words by taking a cosine between the common letters between the dictionary word and the misspelled word. This way we create all the combination of words that are close to the misspelled word by setting a threshold to the cosine similarity and identifying all the words above the set threshold as possible replacement words.

Note: There are several other available techniques similar to cosine similarity, such as *Levenshtein distance*.

Identifying Spelling Errors

Identifying a spelling error is a critical task, and we must keep in mind following points before concluding that a word is misspelled.

Words must not be Proper Noun

While evaluating if a word is misspelled or not, we need to make sure it is not a proper noun. We can do this by checking if the first letter of the word is capital or not. Also, it should not be present in the dictionary.

Words must not be Abbreviation

If we get all capital letters in a word, it is considered as an abbreviation, and hence we can drop it.

Markov Models

In Natural language processing, Markov model is a powerful tool used in identification of misspelled words.

The specialty of Markov model is that it is a state based model and transition from one state to another depends upon the previous state only.

To perform a context based spelling error correction, we need a context based dictionary. This dictionary contains the data related to the topic around which the content is written. We call it as a **Reference dictionary**.

This reference dictionary is used to prebuilt a Markov model, which helps us to perform a strategic walk and identify the probability of the transitions.

Initially, we create all the possible combinations of the text by replacing the misspelled word with the Cartesian set. Then we perform a strategic walk on all the combinations and try to identify the maximum probabilistic pattern.

The maximum probabilistic pattern represents the correct word that should replace the misspelled word.

For example, a context dictionary will look like figure 1.

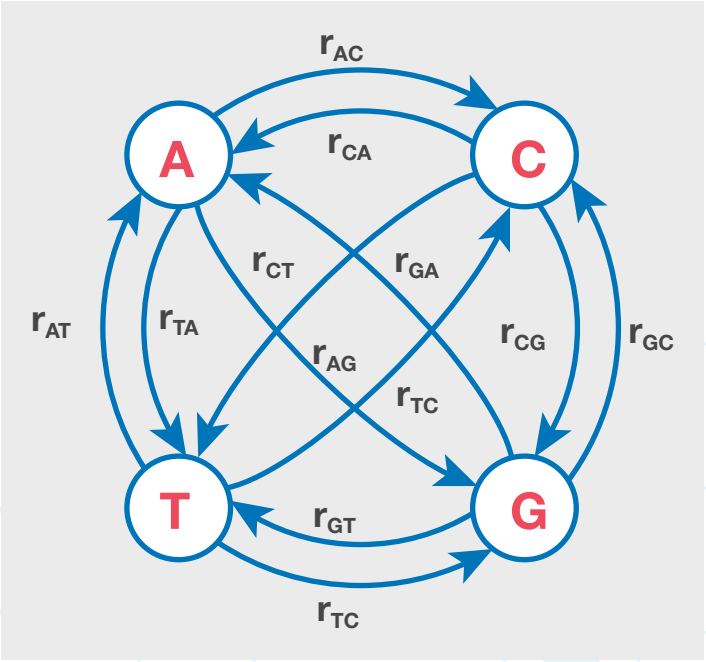


Figure 1

Where A, C, T, G are the words in the reference dictionary.

Deep learning for error correction

Neural networks have evolved magnificently in recent years. as the term deep learning implies that the number of hidden layers and number of nodes have increased drastically. This enables us to learn the sequences and patterns, and associations.

For spelling error correction, we need a context based dictionary that is used to train a deep learning model. This model learns the association of different words with each other and the pattern of occurrence. Using this approach, each word will be converted to a vector, and closely related words will have vectors which are close in distance.

Once we create the Cartesian set, we replace the words with the misspelled word and pass the set to word2vector deep learning engine. We then try to identify the vector distance between the context words, and the replacement word which is closest to the vector of nearby words is selected as the final output.

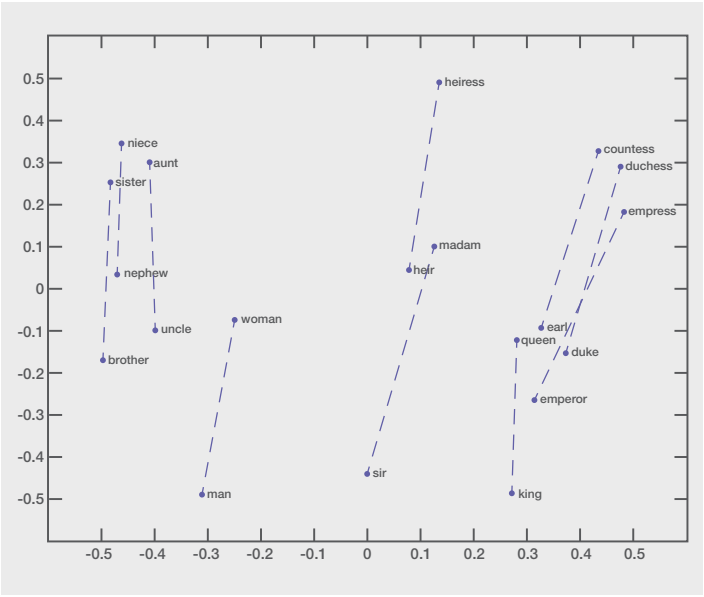


Figure 2

Figure 2 shows the association of related words having nearby vector representation. They are presented in 2-d space.

Advantages of Deep learning:

a) The previous models were based upon transition probability to get best solution. But deep learning is pattern based.

b) Previous models were state based, thus they missed the different structural combinations which can exist, which may not be presented by reference dictionary. Deep learning is vector based approach, and hence doesn't face any state specific problems.

Conclusion

Spelling error correction is a mandatory task that should be performed before any further processing of documents. Thus, understanding the context of the document is an essential part that needs to be taken care of, while replacing the erroneous word.

Ashutosh Vyas

Assistant Manager, Mphasis NEXTlabs

Mr. Ashutosh Vyas is an Assistant Manager at Mphasis Nextlabs. He completed his B.Tech from SKIT, Jaipur Rajasthan Technical University, followed by M.Tech from IIITB Bangalore in Information Technology.He pursued his interest in analytics and machine learning by joining NEXTlabs as a Sr. Analyst, where he worked on different analytical and technical aspects and designed an Agent based simulation model for Mphasis proprietary product – InfraGraf. He designed a machine learning based information retrieval algorithm from documents and sentiment analysis module using deep learning. Ashutosh also published a paper on Life Event Detection in Innovation Asia 2016 conference.

His technical interest resides in machine learning, agent based modelling, optimization theory, graph theory, stochastic based analysis, and time series analytics.

About Mphasis

Mphasis (BSE: 526299; NSE: MPHASIS) applies next-generation technology to help enterprises transform businesses globally. Customer centricity is foundational to Mphasis and is reflected in the Mphasis' Front2Back™ Transformation approach. Front2Back™ uses the exponential power of cloud and cognitive to provide hyper-personalized ($C = X2C^{2}_{TM} = 1$) digital experience to clients and their end customers. Mphasis' Service Transformation approach helps 'shrink the core' through the application of digital technologies across legacy environments within an enterprise, enabling businesses to stay ahead in a changing world. Mphasis' core reference architectures and tools, speed and innovation with domain expertise and specialization are key to building strong relationships with marquee clients. To know more, please visit www.mphasis.com