CS535 - Natural Language Processing Assignment 1
Sentence Segmentation in Urdu

Umair Ahmad
21i-2081

1 Abstract

Preprocessing is the initial phase of training data on any artificial intelligence learning model. The outcomes of preprocessing on English was explored in very comprehensive manners, and the results of it, that how it fulfills the criteria for the training models. However, there is still a lot of work needed to get established for the Urdu language due to the diverse nature of its characters. Space omission and insertion is a huge challenge in preprocessing because of the Urdu language's structures of sentences, before solving these challenges we first need to identify the sentence boundaries which is also known as sentence segmentation or sentence tokenization. Stop word removal, space omission, space insertion, lemmatization, and stemming will be analyzed in the future. However, the aim of this project is to identify the sentence boundaries, Understand the context or sentence building of Urdu, detect the pattern in Urdu sentence, and how we are going to measure the accuracy of the sentence tokenizer.

2 Introduction

Sentence Segmentation is the method of recognizing word borders in Urdu text documents or strings of Urdu text. It identifies the sentence endings and divides the Urdu text into its integral words. It is an initial phase for all language processing systems (LPS), e.g., spell checker, grammar checker, machine translation, information extraction, information retrieval, and part of speech tagging. All LPS required Urdu text with certain word boundaries. Sentence Segmentation disambiguation is the procedure of finding sentence ending, punctuations and related words in the written text which are necessary to complete the sentences and It splits the Urdu text into its sentences.

3 Problem

- Understand the context or sentence building of Urdu.
- Detect the ending of sentences.
- Detect the pattern in Urdu sentence.
- How we are going to measure the accuracy of sentence tokenizer.

4 Solution / Technique for Sentence Segmentation

I named this technique to "My Sentence Tokenizer", First of all, it is important to understand and identify the Urdu sentence building and Urdu grammar in order to solve this specific problem related to sentence segmentation. By thoroughly analyzing the Urdu morphology we will understand that all the tenses have some pattern of sentence finishing, I.e.
Every sentence ends with some specific word

["ھے","ھیں","ھو","ھوں","تھیں","تھے","تھی","تھا","گا","گے","گی","سکا","سکی","ھوگا","ھوگی","گیا","گئیں","کیں"]
Sentence building can be seen here:
https://urdunotes.com/lesson/present-indefinite-tense-in-urdu/ Here we got a pattern that most of the time Urdu sentence end with words mentioned above, we can get those words by manual and auto with two approaches.

- Find the words thoroughly from Urdu grammar.
- Use python NLTK function concordance.

Remember it's not enough to just split the sentence on the basics of sentence-ending words, there are still some rules which most languages support Urdu is also one of them which we called conjunctions. If an ending word is attached with a conjunction word, then this sentence isn't ended yet. I.e.

Sentence:[1]. یہی اصول تاریخ اور انسانی معاشرت پر بھی لاگو ھوتا ھے کہ جس چیز کو جتنی سختی سے دبایا جائے، وہ اتنی ھی قوت سے ابھر کر سامنے آتی ھے چنانچہ اورنگزیب کے بعد بھی یہی کچھ ھوا اور محمد شاہ کے دور میں وہ تمام فنون پوری آب و تاب سے سامنے آ گئے

This is a one sentence we found many ending words like "ھے" and "ھوا " but Urdu did not end the sentence because these words are connected by conjunction words. So, here we got another rule a sentence is not ended if it is connected by a conjunction word. Here are some commonly used conjunction words
["کہ","پر","اور","یا","پہلے","لیکن","کیونکہ","اگر","جب,تک","جب","اسی لیے","ورنہ","یا کہ","یا","نہ","چونکہ","ا
گرچہ","پھر","جونہی","نہ","صرف","بلکہ","تا کہ","تو","جتنا","اتنا","ھم","تا","جو","یہی","جسے","کر","چنانچہ","مگر","جوکہ","کے","جن","کیا","ھی"]

we can get those words by manual and auto with two approaches.

- Find the words thoroughly from Urdu grammar.
- Use python NLTK function concordance.

We will split the words by ending words and check if this is not connected with conjunction and we split it into sentence segmentation. There are also few more details regarding the splitting of sentences, we first need to clean the document text remove the extra spaces, extra punctuation marks. We will also split it on the character like "?", "!", "-". We also need to include numeric and English words.

5 Accuracy Measure

In order to measure the accuracy of My Sentence Tokenizer, we need to apply the human analytical skills to design the test cases for different possible errors which are expected from any sentence tokenizer.
There are popular available Sentence Tokenizer (I.e. Urdu Hack Sentence Tokenizer) we can compare the My Sentence Tokenizer model with them and see how it is superior to the prior sentence tokenizers.
But before that, we will measure it on these test cases mentioned below.

- Check Segmentation Count
- Check White Spaces
- Check Question Mark
- Check Exclamation Mark
- Check English Words
- Check Numeric Digits
- Check relation with Conjunctions
- Check If It Breaks Small Sentences

As we can clearly see in the results available in the given notebook of ipython it passes all the mentioned tests.

6 Comparison of My Sentence Tokenizer with Urdu Hack

As we can clearly see with two example's results that Urdu hack ignore the conjunction words which cannot be tokenized but the Urdu Hack forcefully split it. However, my Sentence Tokenizer can handle these cases and also produce the same results as Urdu Hack's Sentence Tokenizer.

Testcase-1: sentence contains punctuations which cannot be tokenized
"انھوں نے کہا: “آپ نے موسیقی قتل کر دی ھے اسے دفنانے جا رھے ھیں"
Urdu hack split it into two sentence on the other hand My Sentence Tokenizer Split it correctly.

Testcase-2: sentence contains conjunction words which cannot be tokenized [1].
" یہی اصول تاریخ اور انسانی معاشرت پر بھی لاگو ھوتا ھے کہ جس چیز کو جتنی سختی سے دبایا جائے، وہ اتنی ھی قوت سے ابھر کر سامنے آتی
ھے چنانچہ اورنگزیب کے بعد بھی یہی کچھ ھوا اور محمد شاہ کے دور میں وہ تمام فنون پوری آب و تاب سے سامنے آ گئے"

Urdu hack split it into two sentence on the other hand My Sentence Tokenizer Split it correctly.
Include references to specific (recent) papers, like

References

[1] https://www.bbc.com/urdu/regional-44031666?cv=1