AI 4001 - Fundamentals of Natural Language Processing

National University of Computer & Emerging Sciences

Muhammad Shaffay
20I-2391
AI-K

Spring 2023
Assignment 1
Text Segmentation in Urdu

Word Segmentation

Working: For word segmentation, I have used an urdu-words dictionary containing 150K words from urduhack. The dictionary is available at the link 'https://github.com/urduhack/urdu-words'. The algorithm picks one to 10 characters from the unsegmented text and finds the equivalent word in the dictionary. The longest possible word that is present in the dictionary is kept. If after 10 combinations, the word is still not found, the algorithm performs backtracking. This way, the algorithm iterates through the entire unsegmented text and performs segmentation.

First I built a forward approach that iterates through the unsegmented string from index zero to length. This approach gave BLEU score of 0.77. Then I built the reverse approach that reverses the given string and starts making character combinations. After each combination, the algorithm reverses the word and searches from the given dictionary. This technique gave the BLEU score of 0.80. Finally, I merged the two approaches in which if a character does not fit in any combination, backtracking is performed. In backtracking the second algorithm is called which checks combinations from the latest index. The below text shows the pseudo-code for Urdu Word Segmenter.

```
def segmenter(test):

    # remove unnecessary symbols
    # parse through a sentence
        # makes words of length range 1 to 10
            # find if the word is present in the dictionary of 150K+ urdu words
            # overwrite if the new long word is found
        # If the word is found
            # append to list
        # If the word is NOT found
            # Perform backtracking

    # return segmented text
```

Evaluation: For evaluating the results, the BLEU score technique has been used. BLEU (Bilingual Evaluation Understudy) is a metric for automatically evaluating machine-translated text. But it can also be used for other text processing tasks to check the similarity between two texts. BLEU score is a number between zero and one that measures the similarity of algorithm-returned' text to a set of high-quality segmented text. The hybrid approach(forward and reversed combined) gave BLEU score of 0.80.

Example: Below is the unsegmented text and algorithm-returned segmented text.

Unsegmented:تجربہکارہندوستانیآفسپنررویچندرنایشوننےآئندہایشیاءکپ2023ءکیغیریقینیقسمتپراپنیرائےکااظہارکیاہےجوپاکستانمیں ہونےجارہاہےاپنےیوٹیوبچینلپرباتکرتےہوئےرویچندرنایشوننےکہاکہاگرپڑوسیملکبھارتایشیاکپ2023ءمینشرکتکرناچاہتاہےتومقامتبدی لکردیناچاہیے

Segmented: تجربہکار ہندوستانی آف سپنر روی چندرن ایشون نے آئندہ ایشیاء کپ کی غیریقینی قسمت پر اپنی رائے کااظہار کیاہے جو پاکستان میں ہونے جارہاہے اپنے یوٹیوب چینل پر بات کرتے ہوئے روی چندرن ایشون نے کہاکہ اگر پڑوسی ملک بھارت ایشیا کپ میں شرکتکرنا چاہتاہے توم قامت بدیل کردینا چاہیے

Sentence Segmentation

Working: For sentence segmentation, I have used a rule-based approach. The algorithm works by tokenizing the given sentence and analyzing each word. Each word is checked with a list containing the most used sentence-ending words. If conjunctive words occur immediately after sentence-ending words the segmentation is delayed. The algorithm also checks if the sentence continues or ends by comparing it with a list containing powerful sentence-ending words that have more precedence than the other words. The minimum sentence length is kept at 5. The pseudocode for the technique is given below.

```
Def segmenter(text):

    # Tokenizing sentence
    # Parse through tokenized words

        # Condition: Check If End of Paragraph
        # Condition: Check If  '\t' or \n  occurs

        # Condition: Check If Sentence ending words occur
            # Condition: Check If Minimum sentence length  >  5
                # Condition: Continued words are not in range (0,3) of this word
                    # Condition: Check If Next word is not conjunctive
                        # Add Dash
    # Return segmented text
```

Example: Below is the unsegmented sentence text and algorithm-returned segmented text.

Unsegmented: وزیر داخلہ رانا ثناء اللہ کا کہنا ہے کہ عمران خا ن نے جیل بھرو تحریک کا اعلان کیا ہے وہ پہلے بھی اس ہتھکنڈے میں ناکام ہوئے عمران خا ن کو معلوم ہی نہیں کہ جیل میں رہنا کتنا مشکل ہے میڈیا رپورٹس کے مطابق رانا ثناء اللہ نے اپنے بیان میں کہا کہ عمران خا ن کا مقصد سیاسی افراتفری ہے وہ اس میں ناکام ہوں گے عمران خان اپنی زندگی کا صرف ایک دن جیل میں رہے

Segmented: وزیر داخلہ رانا ثناء اللہ کا کہنا ہے کہ عمران خا ن نے جیل بھرو تحریک کا اعلان کیا ہے۔ وہ پہلے بھی اس ہتھکنڈے میں ناکام ہوئے۔ عمران خا ن کو معلوم ہی نہیں کہ جیل میں رہنا کتنا مشکل ہے۔ میڈیا رپورٹس کے مطابق رانا ثناء اللہ نے اپنے بیان میں کہا کہ عمران خا ن کا مقصد سیاسی افراتفری ہے۔ وہ اس میں ناکام ہوں گے۔ عمران خان اپنی زندگی کا صرف ایک دن جیل میں رہے۔