Manha Zamir 17L-4096

Urdu Sentence Segmentation

Introduction:

Sentence segmentation or sentence boundary disambiguation is one of the most crucial part in NLP, where a corpus of text is separated based on the sentences. Other languages such as English, Spanish, etc. use a reasonable approximation of delimiter (full-stop/comma) or case discrimination helps in detecting a sentence boundary. For language such as Urdu, the absence of capitalization makes it a hurdle in indicating a new sentence.

Algorithm:

For this reason, I've implemented a hybrid approach using special character delimiter, end-words and conjunctions for every sentence to identify the end of the sentence. This provides a good accuracy of 81.4%. The algorithm uses a training data and a labelled data. The code works on the following steps:

1) Splitting the Urdu corpus sentences (labeled data) based on delimiters such as '۔' or '؟' using the **split_using_delimiter()** function. This function is similar to the Urduhack's, _split_and_keep() function.
2) Retrieving the end-words and joining words using the **get_endwords_and_conjunctions()** function.
    a) Initializing a list of end-words and conjunction words.
    b) If labeled sentence ends with a delimiter, choose the word before the delimiter as an end-word. Append to the list of endwords.
    c) While the labeled data has not reached the end, if a word is an endword and word+1 is not an endword, then word+1 is a conjunction. Append to the list of conjunctions.
3) Segmenting the training data (unlabeled) using the end-words and conjunction words.
    a) If word is not an end-word, then not the end of sentence.
    b) If word is an end-word but is followed by a conjunction (word+1), then not the end of sentence.
    c) If word is an end-word but is followed by another end-word(s), choose the latter as the end of sentence.

Accuracy:

Using the **acc()** function, we compared the results of our algorithm and labeled data. Urduhack's algorithm showed the accuracy of 74.0%. Compared to this, the algorithm showed an improved accuracy with 81.4%.

Conclusion/Issues for margin in accuracy:

Urdu segmentation can be improved if the following issues are met:
1. Use of special character delimiter in between the sentences (for example; to define ranges).

2. Other punctuations used to define an expression within a sentence (for example; واہ! which could either be a sentence terminator or a part of expression)