Tokenization Practical - 2

Ashish Patidar

The maxmatch algorithm starts from the smallest word or letter in the sentence and keeps on merging words till the longest word is found in the dictionary.

The UD-Japanese_GSD train and test files are used to test the working of the maxmatch and it works just fine. For eg: sentence: "これに不快感を示す住民はいましたが,現在,表立って反対や抗議の声を挙げている住民はいないようです。", is divided into following tokens:

- これ
- に
- 不快
- 感
- を
- 示す
- 住民
- はい
- まし
- たが
- ,
- 現在
- ,
- 表
- 立っ
- て
- 反対
- や
- 抗議
- の
- 声
- を
- 挙げ
- てい
- る
- 住民
- はい
- ない
- よう
- です
- 。