Text Segmentation

Sarvesh , Suchitha , Tarun

Text segmentation is the process of dividing written text into meaningful units, such as words, sentences, or topics. The term applies both to mental processes used by humans when reading text, and to artificial processes implemented in computers, which are the subject of natural language processing. The problem is non-trivial, because while some written languages have explicit word boundary markers, such as the word spaces of written English and the distinctive initial, medial and final letter shapes of Arabic, such signals are sometimes ambiguous and not present in all written languages.

Different problems in segmentation

Word segmentation

Word boundaries

Word segmentation is the problem of dividing a string of written language into its component words.

Intent segmentation

Intent segmentation is the problem of dividing written words into keyphrases (2 or more group of words).

Sentence segmentation

Sentence segmentation is the problem of dividing a string of written language into its component sentences. In English and some other languages, using punctuation, particularly the full stop/period character is a reasonable approximation.

Topic segmentation

Topic analysis consists of two main tasks: topic identification and text segmentation. While the first is a simple classification of a specific text, the latter case implies that a document may contain multiple topics, and the task of computerized text segmentation may be to discover these topics automatically and segment the text accordingly. The topic boundaries may be apparent from section titles and paragraphs. In other cases, one needs to use techniques similar to those used in document classification.