Domain-independent Text Segmentation

Mona Fadaviardakani

THE UNIVERSITY
OF BRITISH COLUMBIA

Motivation

- the problem of separating the documents into coherent segments based on their semantic similarity
- Better representation of the document’s structure can be pre-requisite for lots of other tasks: automatic summarization, Question-Answering, Discourse-analysis and etc.

Related Works

- Supervised and Unsupervised Methods:
  - Unsupervised methods mostly need huge memory, long run-time, and can not generalize well across different text structures and writing styles
  - Supervised methods often require domain expertise and feature engineering which is costly in terms of data annotations
  - Little neural works

Benchmark Model: Attention-based Neural Text Segmentation

Task: A binary classification problem

Input= A samples (sentence, or paragraph, or chapter)

Output=Yes/No tags define whether the sample starts the segment or not

Formal specification:

- Given the document and with respect to **i-th sentence and k the context size**
- We consider K sentences before **i-th (left context)** and K sentences after **i-th (right context)**
- Predict whether the sentence **i-th** denotes the beginning of a new text segment or not

Overview of the Benchmark Model

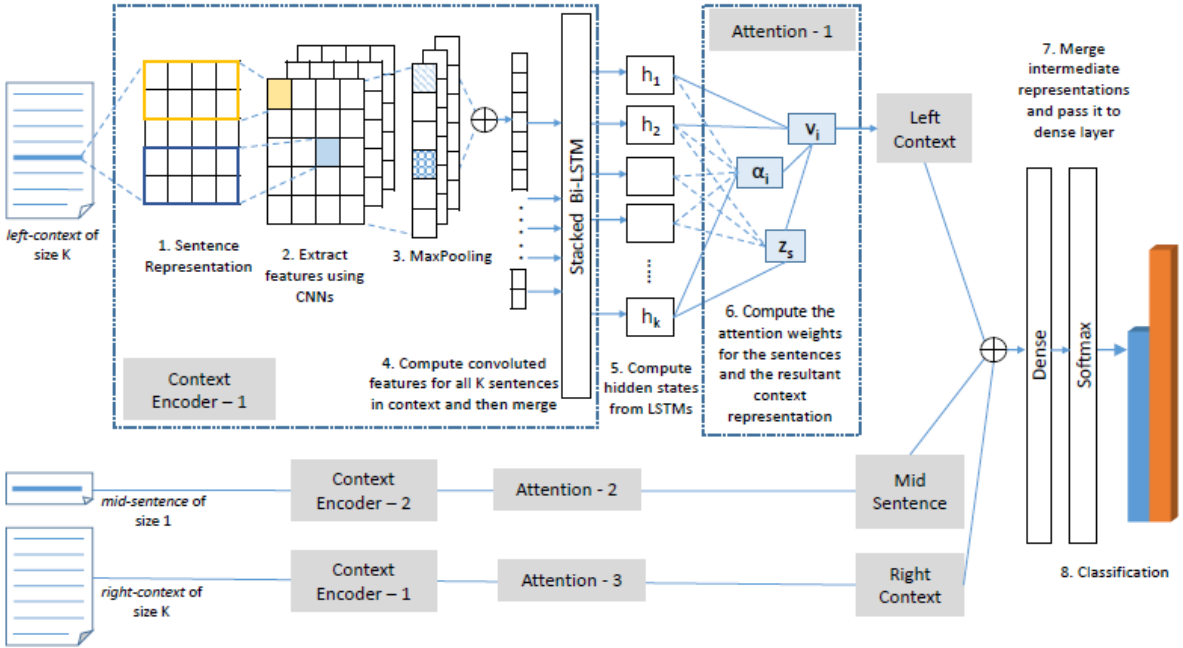


Figure from “Attention-based Nueral Text Segmentation” by Badjatiya et al. 2018

Limitation of the benchmark approach

Challenges=

- Not consider other segment positions in its decision
- Only limits itself to k as the context size and not consider broader range of sentence

Another Approach

Define task as Sequence Labeling Classification Problem

Using Seq2Seq model

- The drawback of these approaches is that the output dictionary is fixed and is not dependent on the input sequence.

Best Approach

Task: Neural Text Segmentation Problem

Input= Sequence of samples (sentence, or paragraph, or chapter)

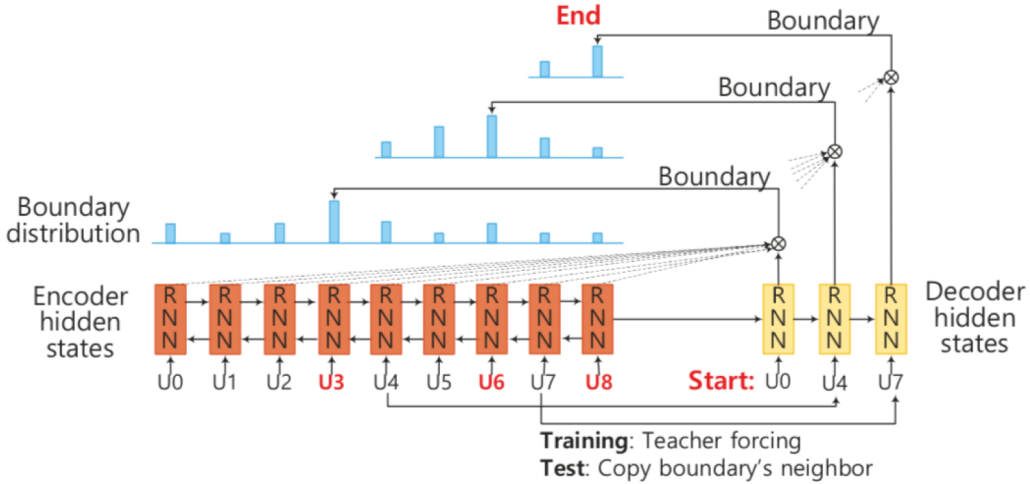
Output= indexes of segment positions in the input sequence

Challenges=

- the issue of variable size output vocabulary.
- Sparsity of the boundaries: capture the dependencies of other segment positions when the boundaries are sparse.

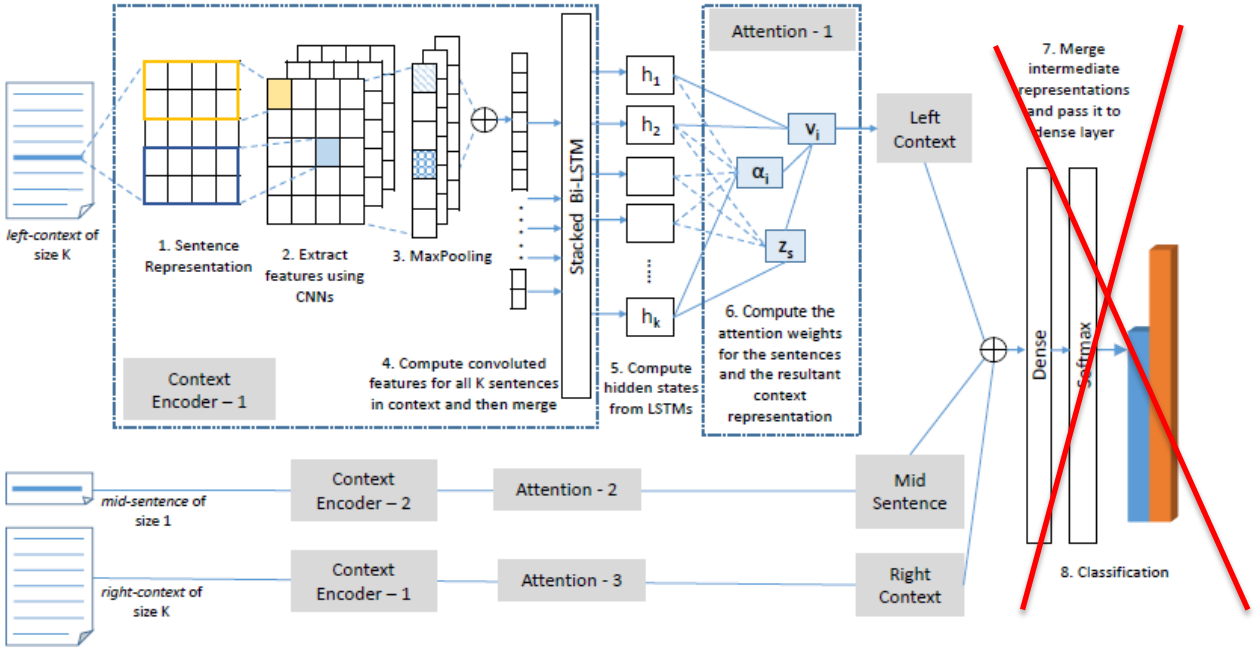
Solution: Multi-layer Pointer Network

- Pointer Networks : New neural architecture to learn the conditional probability of an output sequence

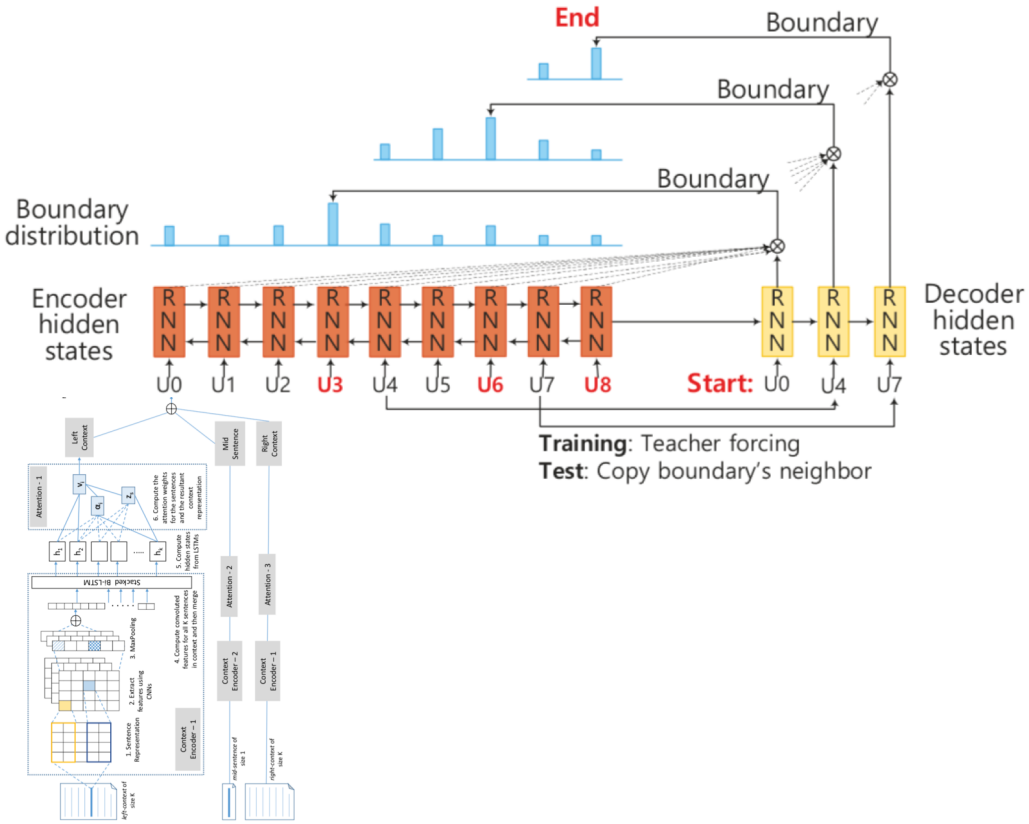


SEGBOT: A Generic Neural Text Segmentation Model with Pointer Network [2018]

My approach: Combination of these two



My approach: Combination of these two



My approach: Expand the Baseline Method

Using the output of concatenation layer of benchmark Model as :

The Rich REPRESENTATION for each sample. → Batch of these representation

↓

Input to the multi-layer pointer network

Two Hyperparameter for this approach:

- The Context Size
- The Pointer Input_Size

Two main advantages over the benchmark:

- Consider broader range of samples
- Consider other samples boundaries positions

Dataset

- standard benchmark datasets :
    - Fiction :Consists of a collection of 85 fiction books downloaded from Project Gutenberg. Segmentation boundaries are the chapter breaks in each of the books.
    - Clinical : Consists of a set of 227 chapters from a medical textbook. Each chapter is marked into sections indicated by the author which forms the segmentation boundaries. It contains a total of 1136 sections
    - Biology: Total BIOGRAPHY data: 11 chapters, 298 paragraphs and 2285 sentences
    - Wikipedia: Consists of randomly selected set of 300 documents having an average segment size of 26

Experiments

- most frequently used measure to evaluate segmentation: PK and WinDiff

Model	Clinical		Fiction		Biography	
Baseline Model	Pk=. 0.318	Windiff: 0.794	Pk=. 0.378	Windiff: 0.308	Pk=. Not report	Windiff: Not report
My approach	0. 630	0.938	0.4789	0.4210	0.3851	0.258

Any Question?