Relevant work

Summaries

[1] SECTOR: A Neural Model for Coherent Topic Segmentation and Classification

Introducing SECTOR – to support machine reading systems – a model that **segments documents into coherent sections and assigns topic labels** to them. In addition, contributes WikiSection, a publicly available data set with 242k labelled sections in English and German from two distinct domains: diseases and cities.

Mentioned are how the tasks of topic modelling, text segmentation and text classification are often done separately; SECTOR unifies them. Topic modelling commonly is about extracting the semantic content of entire documents, not segments – mentioned are LDA models (towardsdatascience example), neural topic embeddings, RNNs. Text segmentation has been done by a wide variety of methods, most recently: LDA, word embeddings, LSTM networks, CNNs. Text classification detects topics on text chunks (paragraphs, sentences) using machine learning methods, such as: SVMs, shallow/deep NNs, CNNs.

- Input →
  - Document $\boldsymbol{D} = \langle \boldsymbol{S}, \boldsymbol{T} \rangle$
    - N sentences $\boldsymbol{S} = [\boldsymbol{s_1}, \ldots, \boldsymbol{s_N}]$
  - Empty segmentation $\boldsymbol{T} = \emptyset$ as input.
- Task →
  - Split **D** into sections $\boldsymbol{T} = [\boldsymbol{T_1}, \ldots, \boldsymbol{T_M}]$
  - $\boldsymbol{T_j} = \langle \boldsymbol{S_j}, \boldsymbol{y_j} \rangle$
    - **S$_j$** is a sequence of coherent sentences $S_j \subseteq S$
    - **y$_j$** describes the common topic for each sentence in **T$_j$**

SECTOR architecture:

1. **Sentence encoding:** transform each plain text sentence into fixed-size sentence vector **x$_k$**, serving as input into the neural network layers. 3 methods:
   - Bag-of-Words Encoding as baseline.
   - Bloom Filter Embedding, which compresses things.
   - Sentence Embeddings, distributional representation based on pre-trained word2vecembeddings.
2. **Topic embedding:** using 2 layers of LSTM (bidirectional, with forget gates) to produce a *"dense distributional representation of latent topics for each sentence"*.
3. **Topic classification:** class labels decoded using one-hot or bag-of-words (for single or multi-labelled sections respectively).
4. **Topic segmentation:** leverage the information encoded in the topic embedding and output layers to segment the document and classify each section.
   - *Newline* embedding as a baseline: split sections at every sentence break and then merge sections that at least share one label in the top-2 predictions.
   - Using deviation of topic embeddings: calculate magnitude of *embedding deviation* (emd) per sentence, peaks of emd are used as starts of a new section.
   - Addition to the previous method: *bidirectional embedding deviation* (bemd).

Results were good; bemd best performance; bloom filters on par with sentence embeddings but is slower.

[2] Text Segmentation as a Supervised Learning Task

Previous work on text segmentation focused on unsupervised methods such as clustering or graph search, due to the lack of quality (real-world & varied) labelled data. In this work, text segmentation is formulated as a supervised learning problem, and presented is a large new dataset, *Wiki-727K*, for text segmentation that is automatically extracted and labelled from Wikipedia.

- Document is **x**, represented as:
  - sequence of N sentences $s_1, \ldots, s_n$.
  - label $y = (y_1, \ldots y_{n\text{-}1})$, represents the segmentation of the document by n-1 binary values, where y_i denotes whether s_i ends a segment.
- Model exists of 2 sub-networks:
  - Lower-level is two-layer bidirectional LSTM, which takes a sentence as input and generates its sentence representation/embedding.
  - Higher-level is two-layer bidirectional LSTM, which takes a sequence of sentence embeddings and tries to predict the segmentation probability for each embedding.

Fairly good results, but the model's performance also depends on the kind of text it tries to predict on (e.g., texts from the chemistry domain are more confusing if it hasn't trained enough with similar texts).

Bibliography

[1] S. Arnold, R. Schneider, P. Cudré-Mauroux, F. A. Gers en A. Löser, „SECTOR: A Neural Model for Coherent Topic Segmentation and Classification," *Transactions of the Association for Computational Linguistics,* pp. 169-184. doi: https://doi.org/10.1162/tacl_a_00261, 2019.

[2] O. Koshorek, A. Cohen, N. Mor, M. Rotman en J. Berant, „Text Segmentation as a Supervised Learning Task," *Proceedings of NAACL-HLT,* pp. 469-473. doi: https://doi.org/10.48550/arXiv.1803.09337, 2018.