Yale

Automatic Generation of Surveys for Topics in NLP

Tomoe Mizutani

[1]Department of Computer Science, Yale University, New Haven, CT

LILY Lab

Introduction

Manually writing a Wikipedia page for every topic can be a labor intensive process, especially for fast-changing topics like AI or NLP. In this work, we explore the task of automatically creating an overview article that serves as a comprehensive summary for a topic of interest. The task involves two stages: 1. creating sections for the survey articles; and 2. creating section-body for the survey articles. This report focus on the section creation part of the project. Specifically, I discuss the different ways of classifying the source text into individual sections with header titles such that the output can be used in the abstractive model.

Some effective methods include clustering with Louvain community detection with Rapid Automatic Keyword Extraction (RAKE). The quality of the output sections can be evaluated using Silhouette score, an intra cluster purity measure. The results of clustering on sample data suggest some promising results for the clustering approaches.

Data

We use a subset of the WikiSum dataset, a collection of over 1.5 million Wikipedia pages and their references. Specifically, we make use of the downloaded version of the corpus that we received from the authors. However, the majority of the data from WikiSum is not suitable for full-document generation, because most pages do not contain many sections to allow for modeling interactions among sections during training. Additionally, much of the section information is found in list or table form. To prepare a suitable dataset, we filter out samples from WikiSum that contain fewer than three sections and those that which contain lists or tables as determined by pre-defined rules.

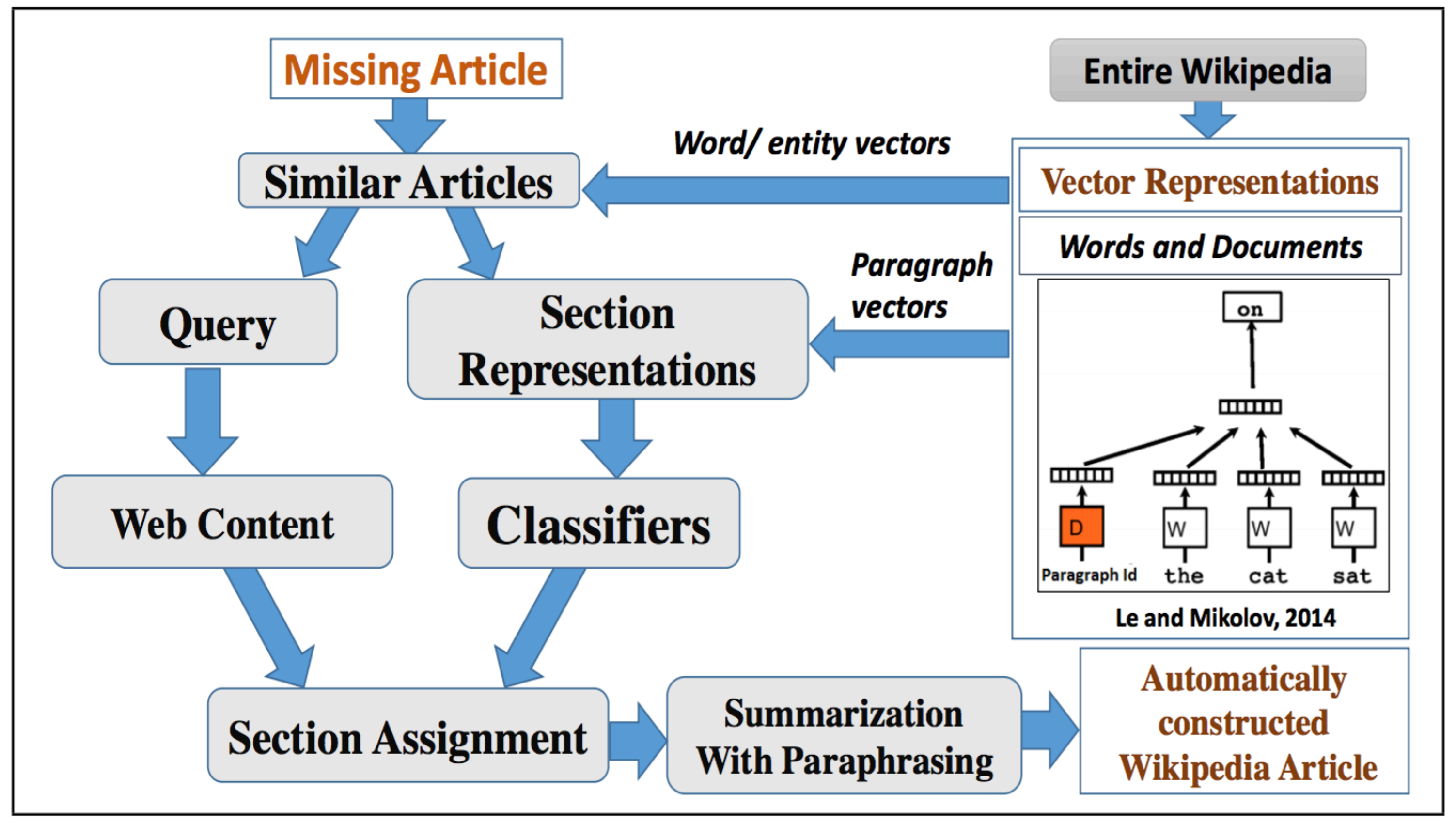

Figure 1. Overview of the survey generation framework proposed in WikiWrite

Target entity: Humboldt Penguins

Cluster	Keywords	Clustered text sample
1	['zavalaga 2001', 'whole blood', 'warming oceans']	The coastline along which the Humboldt penguin is found is particularly susceptible to the influences of El Niño Southern Oscillation ENSO event…
2	['wild places', 'species support', 'siginificant percentatge']	Colonies in Peru and Chile are monitored regularly…
3	['young penguin', 'yearly cycle', 'would achieve']	Egg-laying can occur at any time of year between March and December…
4	['whitish underparts', 'white stripe', 'slightly smaller']	The Humboldt penguin is similar in size to Magellanic penguins , having an average length of Egg-laying can occur at any time of year between March and December…
5	['wild populations', 'zavalaga 2001', 'trachurus murphyi']	The Humboldt 's penguin is rated vulnerable and populations do not exceed 12 ' 000 individuals

Figure 2. Sample output of Louvain clustering and RAKE

Approach

Previous models including WikiWrite involve extracting section headers then filling in those sections. Figure 1 shows a framework that is used in WkiWrite. We consider a bottom-up approach, in which we first split the input text into sections before extracting keywords for each section. This approach allows to extract input-dependent headers. We experiment with two methods of clustering: text segmentation and text clustering.

Segmentation: We experiment with semantic text segmentation with embeddings. This algorithm uses GloVe embeddings and greedy sequence segmentation to semantically segment a text document into any number of k segments. However, if the input data is a series of paragraphs that do not retain their original order rather than a continuous single document, this segmentation method is ill suited for the task.

Clustering: We look into Louvain community to cluster the input document. For sentence representation, I used sentence embeddings with BERT. In Louvain community detection, we construct a graph such that each node represents a sentence in the input document. Then, we compute the cosine similarity between each pair of sentence embeddings. If the similarity value crosses a certain threshold, we add an edge between the two nodes representing the sentences. The algorithm chooses an optimal k number of communities, or clusters. An intra-cluster purity measure can be used to evaluate the effectiveness of clustering. After segmenting or clustering the input into k sections, I extracted keywords from each section using RAKE (Rapid Automatic Keyword Extraction). This domain independent keyword extraction algorithm extracts key phrases in a body of text by analyzing the frequency of word appearance and its co-occurrence with other words in the text. The algorithm could be parametrized to specify the output length.

Evaluation

In the cluster-then-extract approach, Silhouette scores can be used to measure the purity of the resulting clusters.

For a data point *i*, its Silhouette score is defined as

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

where *a(i)* is the mean distance between *i* and all other data points in the same cluster, and *b(i)* is the smallest mean distance of *i* to all points in any other cluster.

I used the *silhouete_score* metrics from scikit-learn to find the mean silhouette score on all samples. By manually inspecting the output clusters on some sample data, I found that high silhouette scores generally correspond to high-quality clusters

Conclusion

We introduced the application of clustering techniques to the extractive portion of the task of automatically generating surveys. The results suggest that Louvain community detection on GloVe embeddings produce quality clusters that can then be used for summarization. The biggest challenge lies in the task of extracting header titles. Although it is often useful to use headers from similar Wikipedia articles, we would also like to be able to generate context-dependent headers using cluster-then-extract approach. Next steps include running the models on larger scale datasets, and running the model end-to-end to observe the effect of the clustering methods on the final summaries.

Acknowledgement

I am deeply grateful for Alex Fabbri for assisting me throughout the project.