SubModular Functions & Text Summarizations

CS 691

R&D Project Report

by

Jayaprakash S

Roll No: 123050045

under the guidance of

Prof. Dr. Pushpak Battacharya

Department of Computer Science and Engineering

Indian Institute of Technology, Bombay

Mumbai

Acknowledgement

I would like to thank my guide, **Prof. Dr. Pushpak Battacharya** for the consistent support and guidance he provided throughout the semester.

Abstract

The aim of this work is to *experiment* with designing a class of submodular functions for document summarization tasks. Designing of submodular functions which scores the summary of text document based on its representatives and divesity in the summary.

List of Figures

Contents

Chapter 1

Introduction

...a wealth of information creates a poverty of attention...

Herbert A. Simon

1.1 Problem Statement

In this project, we are trying to extract keywords from a given news article. Intent of doing keywords extraction is to retrieve an relavant image for a given news article without annotated image. So the extracted keywords should closely match with descriptive metadata of an relevant image. After extracting keywords from text, keywords will be used for retrieving an image using *Image Search System or Engine*.

1.2 Motivation

1.3 Submodular Functions

We are given a set of objects $V = \{v_1, ..., v_n\}$ and a function $F : 2^V \mapsto R$ that returns a real value for any subset $S \subseteq V$.

Subset Function (F)

$F : 2^V \mapsto R$

In text summarization perspective, we are interested in finding subset of bounded size $|S| \leqslant K$ that maximizes the function F.

Maximize the Subset Function

$S = \mathrm{argmax}_{S \subseteq V} F(S)$

Subject to $|S| \leqslant K$.

Finding a subset that maximizes this function is hopelessly intractable. The submodular functions have wide applications in various domain including NLP such text summarization and word alignment.

Example

F might correspond to the value or coverage of a set of sensor locations in an environment, and the goal is to find the best locations for a fixed number of sensors.

if the function is monotone submodular, still the objective is NP-Complete. But there exist a greedy algorithm which will give the near optimal solution.

1.4 Greedy Algorithm

Algorithm 1 : Greedy Algorithm

$G \leftarrow \emptyset$

$U \leftarrow V$

while $U \neq \emptyset$ do

$k \leftarrow \mathrm{argmax}_{l \in U} \frac{f(G \cup \{l\}) - f(G)}{(c_l)^r}$

$G \leftarrow G \cup k$ if $\sum_{i in G} c_i + c_k \leqslant B$ and $f(G \cup \{k\}) - f(G) \geqslant 0$

$U = U\ k$

end while

$v^* \leftarrow \mathrm{argmax}_{v in V, c_v \leqslant B} f(v)$

return $G_f = \mathrm{argmax}_{S \in \{\{v^*\}, G\}} f(S)$

1.5 Proof of Near Optimal Solution

TODO:update later

1.6 Contributions

Contibutions towards this R&D project includes,

1. implementation submodular functions for text summarization using similarity score as TFxIDF.

2. experiments with semantic similarity measures instead of TFxIDF.

3. Using different clustering methods K-means and Single link to improve the diversity of summary sentences.

Chapter 2

Experiments

Good summary of text document expected to have good coverage and non-redundancy (novelty).

Objective functions for extracive summarization usually measure these two components separately and combine them together with tradeoff between encouraging the relavency and penalizing for redundancy.

Objective

$F(S) = L(S) + \lambda R(S)$

F(S) measures the coverage

R(S) rewards diversity

$\lambda > 0$ is a trade-off coefficent.

2.1 Coverage Functions

2.1.1 TFxIDF

Documents were pre-processed by segmenting sentences and stemming words using the Porter Stemmer. Each sentence was represented using a bag-of-terms vector, where we used context terms up to bi-grams. Similarity between sentence i and sentence j, was computed using cosine similarity,

Sentence Similarity (TFxIDF)

$$w_{i,j} = \frac{\sum_{wins_i} tf_{w,i} x tf_{w,j} x idf_w^2}{\sqrt{\sum_{wins_i} tf_{w,i}^2 idf_i^2} \sqrt{\sum_{wins_j} tf_{w,j}^2 idf_j^2}}$$

where $tf_{w,i}$ and $tf_{w,j}$ are the numbers of times that w appears in s_i and sentence

s_j respectively, and idf_w is the inverse document frequency (IDF) of term w, which was calculated as the logarithm of the ratio of the number of articles that w appears over the total number of all articles in the document cluster.

2.1.2 Semantic Measures

2.2 Diversity or Reward Functions

Instead of penalizing the redundancy by subtracting from the objective, rewarding diversity is used here.

Diversity Function

$R(S) = \sum_{i=1}^{K} \sqrt{\sum_{j \in P_i \cap S} r_j}$

P_i is the set of sentences in Cluster i

r_j is the reward for the sentence j

where P_i is a partition of the ground set V into separate clusters. The value r_i estimates the importance of i to the summary. The function $R(S)$ rewards diversity in that there is usually more benefit to selecting a sentence from a cluster not yet having one of its elements already chosen. After a sentence is selected from a cluster, other sentences from the same cluster start having diminishing gain, because of the square root function.

2.2.1 Clustering

1. K-means Clustering

2. Single Link Clustering

2.3 Results

2.4 Summary

In this chapter, we described the experiments done in unsupevised and supervised settings. First unsupervised approach is based on *number of occurences and proximity*. Second unsupervised approach is based on *modified textrank which includes co-referencing for constructing a text graph*.

In the next chapter, we conclude this report with future plans.

Chapter 3

Conclusion and Future work

3.0.1 Conclusion

The problem of retrieving an image that matches the semantics of a text document is difficult. We are trying to solve this problem by extracting important keywords from the text document and using keywords for retrieving a relevant image.

In this report, we discussed our overall system and existing work on keyword extraction based on unsupervised and supervised approaches. Our experiments on keyword extraction based on *Counting and Proximity*, *Modified TextRank*, *Naive Bayes* and *HMM* are discussed with results.

References

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.(JAIR)*, 22(1):457–479.

Lin, H. and Bilmes, J. (2010). Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 912–920. Association for Computational Linguistics.

Lin, H. and Bilmes, J. (2011). A class of submodular functions for document summarization. In *ACL*, pages 510–520.