# Summary Report (10/02/2014)

## Overview

In this week, I have done the following things:

* Got summaries for the new data for both CS2001 and CS2610
* Fixed the distance matrix bug
* Tried different setting of K
* Tried different distance matrices

## CourseMIRROR Summary

### Research Questions:

1. Which representation is the “best” way for summary?

* Words (Unigram)
* Phrases (NP-Soft)
* Sentences (Mead or Unigram-based Extractive summarization)

Note: This is the hardest question to answer. The problem is that I want to know which “representation” is the best way for the summary. However, I just choose a method for each of them. Thus, even if one of the method wins, I cannot conclude this is the best way to represent the summary. Thus, how to design a user study to address this problem?

2. Which summary is better, Unigram or Phrases?

Peer-to-Peer evaluation is easier. I can use the same question set asked in Wenting’s user study.

3. Which summary is better, Mead or Unigram-based Summary?

It is similar to the second question.

### Suggestions by Jingtao

* It’s better to see the complete set of phrases, not just a summary (Is it a good idea?)
* For the Phrases, it is better if we can also list the “frequency”
* To address the question “whether the instructor covers the most important things during the lecture” by using context materials.
  + Identify whether a phrase appears in the lecture (slides)

## Improving the K-Medoid

### Fix the distance matrix

The output of SEMILAR is the similarity of phrases (ranges from 0 ~ 1): the less the value, the less similar between the phrases. Instead, for clustering, the input is a distance matrix. For distance, the less the value, the more similar between two phrases.

Therefore, I changed the similarity matrix (S) to distance matrix (D) by

E is the unit matrix, in other words, a matrix with ones.

Note: the sub bullets (a, b, c) are the possible choices we can try.

The difference of this algorithm and the old one is that the frequency is directly considered by the clustering step (all phrases are passed to the cluster step without compute the Soft-Frequency like before)

### Results



### Observation

Although fixing the Distance matrix largely improve the performance, it is still not better than K-Means. It might due to two reasons:

* The clustering doesn’t consider the frequency of phrases, but just the similarity [Will address it next week]
* The distance metric is just “LexiconOverlap”, maybe other metrics work better.

## Different setting of K

In the paper [2], it introduced a method to set the number of clusters, K, as followings.

V is the number of sentences in the document. In our task, V is the number of phrases.

To test the impact of different setting of K, it could vary from 1 to V, by the following,

In the experiment, I set as the same as [2].

## K-Medoid with more metrics

In the experiment, I also test different distance matrices introduced by SEMILAR (refer to the report 08-28-2014 for more detail).

### Results

When , it means that



Best performances with the optimal parameters:

The last row is the best performances

### Observations:

* With different setting of K, the performances did change a lot
* No method with certain K wins for all the metrics (That’s the problem)
* Even with the best K and the best method (It is optimistic), it cannot beat the previous model (such as R1)
* It is strange that the best R2 for POI is when (no clustering is performed)
  + In this case, there is no advantage from clustering

## TODO

* Send the Summary to the CourseMIRROR Server
* Word/Phrase Cloud Generation
* Different method to extract the keyphrases

## Paper I read

[1] Marujo, L., Viveiros, M., & Neto, J. P. da S. (2013). Keyphrase Cloud Generation of Broadcast News. Information Retrieval. Retrieved from <http://arxiv.org/abs/1306.4606>

<http://www.mendeley.com/share/document/invite/42be582725/?utm_medium=email&utm_source=transactional&utm_campaign=share%2Finvitation-document>

Reference

[2] Wan, X., & Yang, J. (2008). Multi-document summarization using cluster-based link analysis. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’08* (p. 299). New York, New York, USA: ACM Press. doi:10.1145/1390334.1390386