# Summary Report (10/23/2014)

## Overview

In this week, I have done the following things:

## How source information is taken into account for multi-document summarization

In my understanding, the basic multi-document summarization framework doesn’t consider the source into account explicitly, see SumBasic [1], Mead [2]. And the coverage (a summary should cover the information from different documents as much as possible) is not the objective function.

### Centroid

This is used in Mead. By using the centroid, “there is no distinction between a word which occurs many times in the same document or the same number of times across several documents.” [1]

It just considers the total number of occurrences \* IDF. However, by doing this, frequency words will have a higher weight.

### Topic Modeling

This is proposed by [1] in the model TopicSum by introducing a document-specific vocabulary distribution.

## Candidate phrase selection with external content

Although the methods based on chunk parser and syntax tree parser offers a simple automatic way to extract candidate phrases, it is noisy and the extracted are not a complete unit that is meaningful.

To solve this problem, it is better to create a dictionary of phrases with domain knowledge. To do it, I annotated candidate phrases from the external content (slides used by the instructor).

To favor a high coverage, this annotation is performed as a deleting task instead of an adding task. In other words, I manual select phrases for each sentence in the slides and delete the parts that are not phrases. In this way, it makes sure that each sentence in the slides has been looked.

In addition, during the annotation, duplicated phrases are allowed because it provides a chance to get the frequency in the external content. More frequent phrases might be more important.

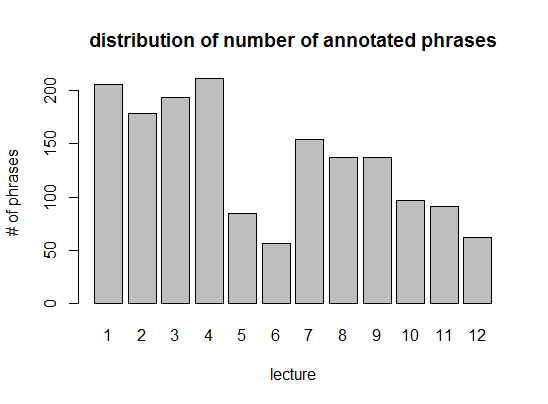
### Preprocessing

The pdf slides are first converted into pure text (images will be discarded) by using the online took Zamzar (http://www.zamzar.com/).

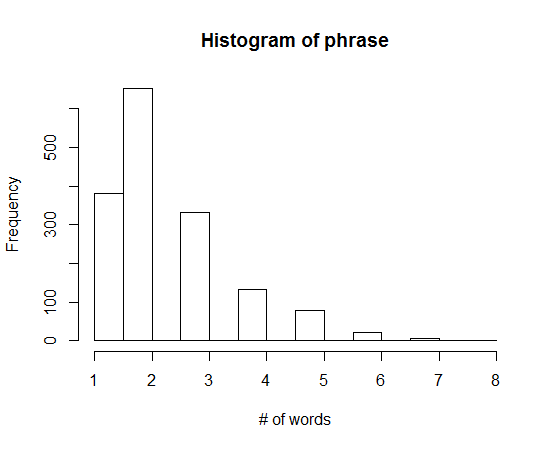
### Annotation

The slides for the first 12 lectures are annotated (Because they have manual summaries).

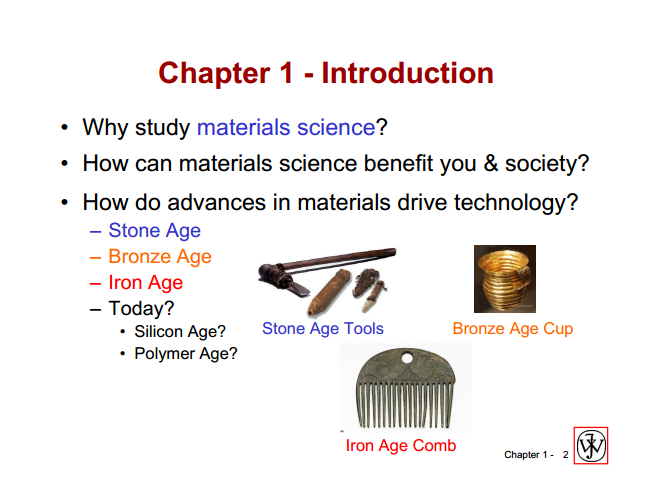
In total, there are 1,596 phrases are extracted.



#### Distribution of length of phrases



An example:



Red rectangles are the annotated phrases.

### Candidate phrase extraction

Since majority of the candidate phrases are less than 5 words. Thus, I will extract the ngrams up to 5 for each sentence in the students’ responses. If a ngram is in the candidate phrase dictionary, it will be selected as a candidate phrase.

#### Overlap handling

One issue to address is to handle the overlap of the phrases. I’m using maximal matching to solve the problem, which is usually used in word segmentation for Asia languages. In other ways, it will select the longest candidate if there is an overlapping.

For example, for the sentence “I’m interested in gas solution or solid solution”, it could match “gas solution or solid solution”, “gas solution” and “solid solution”. However, I will only output the longest one “gas solution or solid solution”.

### Results based on phrase candidate

TODO

## Different clustering methods

TODO

## Paper I read

[1] Haghighi, A., & Vanderwende, L. (2009). Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 362–370). Boulder, Colorado: Association for Computational Linguistics.

<http://www.mendeley.com/share/document/invite/4afc022306/?utm_medium=email&utm_source=transactional&utm_campaign=share%2Finvitation-document>

## Reference

[2] Radev, D. R., Jing, H., Styś, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. *Information Processing & Management*, *40*(6), 919–938.