# Student Reflects Summarization

We leveraged two types of summarization techniques to generate the summaries of the reflective feedback from the students: extractive summarization and shallow summarization. The former one outputs important sentences and the latter outputs important key words and phrases.

## Extractive summarization

The public available summarization toolkit Mead ([www.summarization.com/mead](http://www.summarization.com/mead)) is a widely used unsupervised extractive summarization techniques.

MEAD will first rank the reflection sentences by a linear combination of three features: Position, Centroid and Length. According to this ranking, the top 3 sentences will be the output summary.

## Shallow Summarization

Shallow summarization is another promising way for student reflects beyond extractive summarization for at least three reasons. First, the limited screen size makes a summary with several sentences are hard to read especially when they are long. However, showing a summary of reflects back to students give them a chance to review peer feedback. Thus, shallow summarization techniques are proposed to address this issue. Second, shallow summarization servers a dense compress of the student reflects which can provide more coverage compared to sentences summary with the same space limit. Third, human summaries of student reflects tend to be phrases and they could be easily understood by the instructors without providing full sentences.

Shallow summarization task definition:

* + Input: a collection of *d* documents with a total number of *n* sentences
  + Output: a list of important key words/phrases. The total number of words in the list is less than R.

The difference between shallow summarization and extractive summarization is the granularity of the output unit: sentences vs. words/phrases.

For this task, we proposed three different approaches as following.

### Unigram Shallow Summarization

The most straightforward idea to do shallow summarization is to output the most frequent words (unigrams). It could be represented as either a list of words or as a word cloud.

### Keyphrase extraction Summarization

The task of automatic keyphrase extraction is defined as “the automatic selection of important and topical phrases from the body of a document” (Hasam et al., 2014). It could serve as a shallow summarization method.

We implemented the shallow summarization over the state-of-art keyphrase extraction methods implemented by Maui (<http://code.google.com/p/maui-indexer/>), which also used in (Marujo et al., 2013).

### Phrase Clustering Summarization

We observe that a considerable number of concepts introduced in student reflects are noun phrases. At the same time, one of the challenge is that these concepts have different forms (i.e. in morphology or synonym).

In addition, a good summarization should have a high coverage of the original content but a low duplication within the summary.

To address this issue, we proposed a clustering approach at a phrase level for shallow summarization by considering the semantic similarity between the phrases. We adopt a clustering paradigm because by doing this, similar concepts will be grouped into one cluster (minimize the duplication) and different concepts will be split into different clusters (maximize the coverage).

The proposed shallow summarization algorithm by phrase clustering is as follows:

1. Select the phrase candidates for each sentence s in every document
2. Get a collection of all the extracted phrases as
3. Cluster the phrases
4. Rank the clusters by the number of phrases in the cluster
5. Select one phrase for each of cluster according to the ranking until reaching the R limit

## Evaluation

ROUGE scores have been widely used to evaluate automated text summarization (i.e. DUC 2004, DUC 2005). It has been shown that ROUGE is very effective to measure document (Lin, 2014).

The ROUGE scores measures summary quality by counting the overlapping units such as n-gram between the candidate summary and the reference summary. It is defined by formula below:

Where stands for the length of n-gram and is the number of n-grams co-occurring in a candidate summary and a set of reference summaries.

We adopt ROUGE-1, ROUGE-2 and ROUGE-SU4[[1]](#footnote-1) in our evaluation metrics. The more the scores, the better the summarization.

The data set we use is introduced in (Menekse et al., 2011).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Point of Interest | | | Muddiest Point | | | Learning Point | | |
|  | R1 | R2 | R-SU4 | R1 | R2 | R-SU4 | R1 | R2 | R-SU4 |
| Mead | 25.07% | 8.33% | 5.43% | 25.03% | 8.51% | 6.23% | 14.47% | 1.72% | 2.28% |
| Unigram | 35.22% | 1.72% | 9.19% | 34.61% | 1.95% | 8.71% | **23.55%** | 0.89% | 4.22% |
| Keyphrase | 22.91% | 6.04% | 4.83% | 23.92% | 8.70% | 5.61% | 16.95% | 2.07% | 2.75% |
| Phrase Clustering | **37.74%** | **13.06%** | **11.93%** | **36.69%** | **10.07%** | **9.61%** | 22.95% | **3.56%** | **4.55%** |

\* For the shallow summarization, we set R=30 for the experiments.

\* For phrase clustering, we used the toolkit Senna (<http://ml.nec-labs.com/senna/>) to extract the noun phrases, K-medoids as clustering algorithm (K = 0.3 \* total number of candidate phrases) and the semantic similarity metric (optimumComparerLSATasa) implemented in SEMILAR (Rus et al., 2013).

For the table above, we can see the phrase clustering method performs very well on all the three reflects.

## Example Output

A manually summary of muddiest point for one lecture:

* Grading process
* Homework assignments
* Differences between types of bonding

Mead Summarization (extractive)

[1] The normalized score is still a little vague and could a bit more discussion about it

[2] I think that even though grade normalizing was interesting, I think it was somewhat confusing and needs a little more explaining

[3] Nothing presented was confusing, material that was given on the pre-test was confusing (Need to learn those topics)

Shallow Summarization

|  |  |  |
| --- | --- | --- |
| **Unigram** | **Keyphrase** | **Phrase Clustering** |
| grading | normalized score | a little |
| confusing | bit more discussion | grading |
| clear | homework is due | the grading scale |
| bonds | due The class | discussion |
| material | structure regarding assignments | bonds |
| scale | class structure regarding | the expectation |
| test | little vague | materials |
| bonding | grading | the tests questions |
| class | review bonds Normalizing | the grading system |
| shape | extrusion and wire | the class |
| homework | need to review | the activity |
| bit |  | the pre- test |
| molten |  | the normalized score |
| assignments |  | what homework |
| activity |  |  |

## Future Work

In future work, we can improve the interface by using phrase cloud (Yatani et al., 2011), which is an extension of word cloud. In addition, we can improve the summarization by using the helpfulness-guided summarization (Xiong 2013).



Figure ?: A phrase cloud interface used in product reviews summarization

# References

Lin, C. (2004). ROUGE : A Package for Automatic Evaluation of Summaries. *In Information Sciences*, *16*, 25–26.

Yatani, K., Novati, M., Trusty, A., & Truong, K. N. (2011). Review spotlight. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI ’11* (p. 1541). New York, New York, USA: ACM Press. doi:10.1145/1978942.1979167

Hasan, K. S., & Ng, V. (2014). Automatic Keyphrase Extraction: A Survey of the State of the Art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1262–1273). Baltimore, Maryland: Association for Computational Linguistics.

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

Rus, V., Lintean, M., Banjade, R., Niraula, N., and Stefanescu, D. (2013). SEMILAR: The Semantic Similarity Toolkit. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, August 4-9, 2013, Sofia, Bulgaria.

1. ROUGE-SU4 is an extension of a skip-bigram version of ROUGE by counting unigrams and skip-bigram by up to four words. [↑](#footnote-ref-1)