# Summary Report (10/16/2014)

## Overview

In this week, I have done the following things:

* Sentence splitting for Maui
* Whether the summary comes from the different students
* Whether the summary comes from high quality responses

## 1. Sentence splitting for Maui

In Maui, the sentences are tokenized by the punctuations and one sentence might have multiple lines.

However, some student reflects have no punctuations in the end of sentence. That’s why some of the keyphrases extracted by Maui do not make sense. To solve it, I added a period to each of the sentence to fit it to Maui.



* It does improve the keyphrase extraction performance. However, it is still not better than the Clustering method.

### An Example

A manually summary of muddiest point for one lecture:

[1] In class activities (Specially the first activity)

[2] Atomic Packing Factor and relation between a&r

[3] Which structure produce which materials' properties and how unit cell affects the processing

Extractive Summarization based on Mead

[1] Whethere we describe unit cell lengths in terms of a or c

[2] Unit cell transformation pictures in 1st activity, 2nd activity (characteristics) # of atomic radii along touch directions with in unice cell

[3] The atomic packing factor idea was very confusing and needed more explanation

Shallow Summarization

|  |  |  |
| --- | --- | --- |
| **Unigram** | **Keyphrase** | **Phrase Clustering** |
| unit | Structure characteristics | how unit cell |
| atomic | atomic packing | our activities |
| cell | packing factor | the material |
| activity | unit cell | which structures |
| confusing | Whethere we describe | structure characteristics |
| factor | unit cell lengths | the math |
| activities | describe unit cell | atomic packing factor |
| processing | lengths in terms | apf |
| packing | describe unit | the properties |
| properties | Caculating structure characteristics | a little more explanation |
| length | atomic | the exact difference |
| apf | slides How unit | different crystal systems |
| structures |  |  |
| characteristics |  |  |
| fcc |  |  |

**\* not all the unigrams are listed here due to space limit**

## 2. Whether the summary comes from the different students

To address this issue, the student id is associated with his/her response during the whole processing. And the output of the summary is a list of triple <T, W, S>

* T is the summary text (word, phrase, sentence)
* W is the weight
* S is the source of the summary (a list of students’ id who give the summary)

To measure the quality of the summary, two metrics are defined according to the source of the summary: diversity and coverage

### Diversity ratio

Diversity is usually used in ecology, to measure how many different types (such as species) there are in a dataset. I used [Shannon index](http://en.wikipedia.org/wiki/Diversity_index) as the metric.

It is defined as:

 H' = -\sum_{i=1}^R p_i \ln p_i 

Where, R is the number of students covered in the summary

is the probability that the i^th student is covered by the summaries, defined as

### Coverage ratio

A good summary should have a high coverage [1]. Thus, for our data set, we can measure the coverage by the following formula

Note: the students without an answer (“none”, “n/a”, “nothing”) are not counted, which is an underestimate because these students could be treated to be covered by the summary.)

### Results

The more the diversity and coverage ratio, the better the summary.

#### Coverage

|  |  |  |  |
| --- | --- | --- | --- |
| Methods | POI | MP | LP |
| unigram | 0.922 | 0.915 | 0.855 |
| Cluster-optimumComparerLSATasa | 0.445 | 0.449 | 0.378 |
| Cluster-npsoft | 0.406 | 0.386 | 0.414 |

#### Diversity

|  |  |  |  |
| --- | --- | --- | --- |
| Methods | POI | MP | LP |
| unigram | 6.684 | 6.491 | 5.964 |
| Cluster-optimumComparerLSATasa | 3.102 | 3.241 | 2.890 |
| Cluster-npsoft | 2.963 | 3.025 | 3.027 |

#### Observation

* Unfortunately, the unigram baseline won on both of the new metrics. The problem is that the normalization by the length of the output. It makes sense the unigram model will win for the coverage model because they are the most frequency words shared by the students. While, for the diversity, the problem is the number of entries in the summary is more than the other two given a word limit of 30.
  + It might be better to design other metric or normalized with another way to get the diversity

## 3. Whether the summary comes from high quality responses

The quality score is defined as:

### Results

The weeks that have no quality annotation are removed (week 1).

The quality of ‘a’ (I understood everything) is treated as ‘0’.

|  |  |
| --- | --- |
| Methods | MP |
| unigram | 1.973 |
| Cluster-optimumComparerLSATasa | 1.997 |
| Cluster-npsoft | 2.018 |

Both of the two models are better than the unigram model.

## Candidate phrases from external content

TODO

## Paper I read

[1] Liu, Z., Li, P., Zheng, Y., & Sun, M. (2009). Clustering to find exemplar terms for keyphrase extraction. *Proceedings of the 2009 Conference on …*. Retrieved from http://dl.acm.org/citation.cfm?id=1699544

http://www.mendeley.com/share/document/invite/5f20826103/?utm\_medium=email&utm\_source=transactional&utm\_campaign=share%2Finvitation-document