# Summary Report (09/25/2014)

## Overview

In this week, I have done the following things:

* Similarities of POI, MP, LP

## Similarities of POI, MP, LP

I get the similarities (JSD, see report 09-18-2014) among them with the following setting:

* Students: Use the students’ responses (Gold standard)
* TA: Use the TA’s summary
* Mead: Summary from the Mead toolkit
* Shallow-Unigram: Shallow summary according to the unigram
* Shallow-NP-Hard: Shallow summary according to the Hard Frequency NP
* Shallow-NP-Soft: Shallow summary according to the Soft Frequency NP
* Shallow-SyntaxNP-Hard: Shallow summary according to the Hard Frequency NP extracted by Syntax Tree
* Shallow-SyntaxNP-Soft: Shallow summary according to the Soft Frequency NP extracted by Syntax Tree
* Shallow-ClusterNP-Soft: Shallow summary according to the Soft Frequency NP with clustering based duplicating removing
* Extractive-TopicS: Extractive summary based on Topic Signature
* Extractive-Unigram: Extractive summary based on Unigram

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | JSD(POI,MP) | JSD(POI,LP) | JSD(MP,LP) | Flag |
| Students | 0.205 | 0.264 | 0.252 | TRUE |
| TA | 0.345 | 0.500 | 0.474 | TRUE |
| Mead | 0.300 | 0.359 | 0.349 | TRUE |
| Shallow-Unigram | 0.393 | 0.482 | 0.497 | FALSE |
| Shallow-NP-Hard | 0.338 | 0.419 | 0.350 | TRUE |
| Shallow-NP-Soft | 0.321 | 0.413 | 0.374 | TRUE |
| Shallow-SyntaxNP-Hard | 0.354 | 0.370 | 0.391 | FALSE |
| Shallow-SyntaxNP-Soft | 0.346 | 0.420 | 0.390 | TRUE |
| Shallow-ClusterNP-Soft | 0.354 | 0.423 | 0.447 | FALSE |
| Extractive-TopicS | 0.333 | 0.359 | 0.368 | FALSE |
| Extractive-Unigram | 0.311 | 0.344 | 0.350 | FALSE |

### Observations

* The TA’s summary doesn’t change the similarities, which shows that TA’s summary is good
* The Mead, Shallow-NP-Hard and Shallow-NP-Soft and Shallow-SyntaxNP-Soft are good.
* Unfortunately, the Clustering based method failed.

## Normalization of the Text

The table below lists all the OOV (out of vocabulary words), according to the Word list (<http://www-01.sil.org/linguistics/wordlists/english/wordlist/wordsEn.txt>).

Among them, I found 4 short forms:

* & -> and
* CW -> cold working
* HW -> homework
* RA -> reduction area

From this list, I don’t think the short forms are a big problem because they are not common. Instead, incorrect spelling might be a bigger problem.

|  |  |  |
| --- | --- | --- |
| OOV | Counts | Short for |
| & | 13 | and |
| s | 8 |  |
| % | 5 |  |
| cw | 5 | cold working |
| 3 | 4 |  |
| hw | 4 | homework |
| 1 | 3 |  |
| ra | 2 | reduction area |
| difusion | 2 |  |
| 4 | 2 |  |
| α | 2 |  |
| β | 2 |  |
| recrystallization | 2 |  |
| r | 2 |  |
| t | 2 |  |
| calulation | 1 |  |
| tas | 1 |  |
| !) | 1 |  |
| petch | 1 |  |
| microstuctures | 1 |  |
| labling | 1 |  |
| behaivior | 1 |  |
| strenghts | 1 |  |
| ), | 1 |  |
| tthermal | 1 |  |
| picures | 1 |  |
| bcc | 1 |  |
| apf | 1 |  |
| griffith | 1 |  |
| elingation | 1 |  |
| differentprocesses | 1 |  |
| indecies | 1 |  |
| # | 1 |  |
| + | 1 |  |
| labled | 1 |  |
| 2 | 1 |  |
| 7 | 1 |  |
| stell | 1 |  |
| differents | 1 |  |
| 8 | 1 |  |
| sturation | 1 |  |
| foremd | 1 |  |
| rlationship | 1 |  |
| fil | 1 |  |
| extremly | 1 |  |
| impac | 1 |  |
| spectially | 1 |  |
| .) | 1 |  |
| dislocastion | 1 |  |
| anealing | 1 |  |
| ductil | 1 |  |
| …) | 1 |  |
| fcc | 1 |  |
| …, | 1 |  |
| ao | 1 |  |
| eutectic | 1 |  |
| sfirst | 1 |  |
| d | 1 |  |
| saturatioed | 1 |  |
| i | 1 |  |
| activitiy | 1 |  |
| tepm | 1 |  |
| fick | 1 |  |
| nw | 1 |  |
| 5th | 1 |  |

## Syntax-NP Clustering

Instead of using the NP extracted with Chunk parser, I’d like to use the NP extracted by the syntax tree, which shows better results for the model “Shallow-SyntaxNP-Soft”.

### Results



### Observations

* Clustering based SyntaxNP model get the best R2 scores among all the three points.

## K-medoids based Summarization

“The k-medoids algorithm is another clustering algorithms related to the k-means algorithm. Both the k-means and k-medoids algorithms are partitional (breaking the dataset up into groups) and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k-means algorithm, k-medoids chooses datapoints as centers ([medoids](http://en.wikipedia.org/wiki/Medoids) or exemplars) and works with an arbitrary matrix of distances between datapoints.” (From Wikipedia, http://en.wikipedia.org/wiki/K-medoids)

As much as I know, this clustering method hasn’t been used in the summarization. However, I think it fits more for the task of extractive summarization than K-means for at least two reasons. First, it selects one of the samples as the center of that cluster. It is very useful because this center could be the one that would be extracted. Second, there is no need to represent the datapoint as a feature vector, it works with any arbitrary distance matrix. It needs only a distance function. Thus, we can use the semantic similarity as the feature function.

### Implementation

1. Initialize: randomly select (without replacement) k of the n data points as the medoids
2. Associate each data point to the closest medoid. ("closest" here is defined using any valid [distance metric](http://en.wikipedia.org/wiki/Metric_space), most commonly [Euclidean distance](http://en.wikipedia.org/wiki/Euclidean_distance), [Manhattan distance](http://en.wikipedia.org/wiki/Manhattan_distance) or[Minkowski distance](http://en.wikipedia.org/wiki/Minkowski_distance))
3. For each medoid m
   1. For each non-medoid data point o
      1. Swap m and o and compute the total cost of the configuration
4. Select the configuration with the lowest cost.
5. Repeat steps 2 to 4 until there is no change in the medoid.

I used the python toolkit [Biopython](http://biopython.org/wiki/Main_Page) to implement this algorithm.

### Results

I used the LexiconOverlap as the semantic distance.



### Observation

* Unfortunately, this model doesn’t work as expect. The problem is that the K-Medoid clustering algorithm will cluster the majority phrases into one single cluster when using the lexiconOverlap similarity

Examples:

|  |  |  |
| --- | --- | --- |
| Index | Phrase | Medoid |
| 0 | various materials | 13 |
| 1 | the grade normalizing | 13 |
| 2 | the most | 13 |
| 3 | grades | 3 |
| 4 | how materials | 4 |
| 5 | the application | 13 |
| 6 | all materials | 13 |
| 7 | the semester | 13 |
| 8 | mse 100 | 8 |
| 9 | a normalized grading scale | 13 |
| 10 | extrusion | 10 |
| 11 | this class | 13 |
| 12 | different part | 12 |
| 13 | tungsten | 13 |
| 14 | class activity | 13 |
| 15 | part | 15 |
| 16 | the glass | 13 |
| 17 | this first class | 13 |
| 18 | the 3 families | 13 |
| 19 | the group activity | 13 |
| 20 | notes | 13 |
| 21 | this semester | 13 |
| 22 | only 3 families | 22 |
| 23 | my own words | 23 |
| 24 | light bulb | 13 |
| 25 | professors " student | 13 |
| 26 | bulb filaments | 26 |
| 27 | 't | 13 |
| 28 | the pre- test | 13 |
| 29 | matching | 13 |
| 30 | what | 13 |
| 31 | the uses | 13 |
| 32 | the bonding test | 13 |
| 33 | 's engineer " page | 13 |
| 34 | the hip thing | 13 |
| 35 | we | 13 |
| 36 | the activity | 13 |
| 37 | the most interesting part | 13 |
| 38 | understanding | 13 |
| 39 | a group | 13 |
| 40 | a bike | 13 |

I will try different semantic metrics later.

## Different setting of K

There are two different settings introduced by the paper [1, 2].

TODO: I will try to implement them.

## Paper I read

[1] 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

[2] Zhang, P., & Li, C. (2009). Automatic text summarization based on sentences clustering and extraction. In *2009 2nd IEEE International Conference on Computer Science and Information Technology* (pp. 167–170). IEEE. doi:10.1109/ICCSIT.2009.5234971

Check out annotation by clicking the link below:

<http://www.mendeley.com/share/document/invite/c7d7138708/?utm_medium=email&utm_source=transactional&utm_campaign=share%2Finvitation-document>