# Summary Report (09/18/2014)

## Overview

In this week, I have done the following things:

* Analysis of TA’s summary
  + Word Frequency
  + POS
  + Phrase

## Analysis of TA’s summary

One key step to get a good summary is to see what a “good” summary is. In our case, I am going to see what the TA’s summary looks like, including

* Length Distribution (done, see the 07-17-2014 report)
* Word Distribution
* POS Distribution
* Phrase Type

### Word Frequency

Hypothesis: Topic-related words in the TA’s summary are not shared among weeks. It is because the students’ responses are based on the lectures and their topics are different.

#### Top 20 frequent words:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| POI | | MP | | LP | |
| of | 19 | of | 19 | activities | 14 |
| and | 14 | and | 15 | group | 12 |
| the | 12 | the | 12 | the | 10 |
| materials | 8 | what | 8 | and | 9 |
| on | 7 | how | 8 | to | 6 |
| how | 6 | activity | 7 | graphs | 6 |
| to | 5 | properties | 7 | are | 5 |
| properties | 5 | & | 7 | class | 5 |
| & | 5 | on | 6 | help | 4 |
| 3 | 4 | grain | 6 | examples | 4 |
| activity | 4 | unit | 6 | pictures | 4 |
| s | 4 | to | 6 | visuals | 3 |
| their | 4 | in | 6 | on | 3 |
| examples | 4 | cw | 5 | of | 3 |
| metal | 4 | between | 5 | they | 3 |
| different | 4 | types | 5 | helps | 3 |
| phase | 4 | % | 5 | hw | 3 |
| a | 4 | for | 4 | board | 3 |
| real | 3 | cell | 4 | more | 3 |

#### Observations:

* The most common shared words between different weeks are functional words like “of, the, and, on, to”, etc. However, there are some terms that appears among different weeks, such as “activities” for LP and MP, “properties” for POI. It means that students did give similar answers for different weeks.
* The TA uses simple forms:
  + CW -> Cold working (a domain phrase)
  + HW -> Homework
  + & -> and
* “and” is commonly used. It means that many of the summary involves parallel phrases.
  + Differents bonds **and** their effect
  + Pictures, diagram **and** examples
  + Atomic Packing Factor **and** relation between a**&**r

#### Similarity among POI, MP, LP (The first time to see how different among the three topics)

Form the “Top 20 frequent words” table, I found POI and MP are more similar with each other compared to LP. Thus, I’d like to see how similar among them using the JSD metric. I used the distribution of the unigrams among the TA’s summary.

Jensen–Shannon divergence (JSD)

{\rm JSD}(P \parallel Q)= \frac{1}{2}D(P \parallel M)+\frac{1}{2}D(Q \parallel M)

Where,

M=\frac{1}{2}(P+Q)

D_{\mathrm{KL}}(P\|Q) = \sum_i \ln\left(\frac{P(i)}{Q(i)}\right) P(i).\!

Note, if Q(i)=0, it implies P(i)=0

#### Results:

The smaller the value, the more similar between them. ‘0’ means “the same”.

|  |  |  |  |
| --- | --- | --- | --- |
|  | POI | MP | LP |
| POI | 0 | 0.345 | 0.500 |
| MP | 0.345 | 0 | 0.474 |
| LP | 0.500 | 0.474 | 0 |

#### Observations:

* POI and MP are indeed more similar than POI to LP, MP to LP

### POS

#### Single POS Distribution

* X is the POS tag, the top three
  + NN -> Noun, singular or mass
  + NNS -> Noun, plural
  + IN -> Preposition or subordinating conjunction

The complete POS tag can be found at

<https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html>

Here are the detail numbers.

|  |  |  |  |
| --- | --- | --- | --- |
| pos | POI | MP | LP |
| NN | 128 | 151 | 46 |
| NNS | 48 | 49 | 42 |
| IN | 39 | 47 | 21 |
| JJ | 27 | 38 | 23 |
| NNP | 14 | 33 | 11 |
| CC | 21 | 23 | 11 |
| , | 14 | 30 | 7 |
| VBG | 12 | 12 | 23 |
| DT | 18 | 16 | 13 |
| -LRB- | 9 | 10 | 6 |
| -RRB- | 9 | 10 | 6 |
| : | 10 | 8 | 6 |
| PRP | 10 | 7 | 6 |
| VBZ | 7 | 10 | 5 |
| VBP | 1 | 9 | 9 |

#### Observations:

* Although the lexicon distributions are very different between POI, MP and LP, the Part-of-Speech distributions are similar with each other.
* The noun words are popular (NN, NNS, NNP). It shed a light that why the NP-based method works well.
* The Adjective words (JJ, JJS) are common, too. However, they are not as popular as noun. It can explain why the Adjective-Noun Phrases method are not better than just the NP phrases.
* The “CC” is common. “and”, “or”, “&” will be tagged as “CC”, which is confirmed by the lexicon unigram distribution.

#### Bigram-POS and Trigram-POS

Unigram-POS (single POS) doesn’t give a sense what’s the structure of a TA’s summary. Thus, I decide to extend it to Bigram-POS, Trigram-POS

#### Most frequent Bigram-POS

|  |  |  |  |
| --- | --- | --- | --- |
| pos | POI | MP | LP |
| NN NN | 26 | 37 | 6 |
| JJ NN | 17 | 25 | 10 |
| NN IN | 19 | 16 | 5 |
| IN NN | 15 | 18 | 0 |
| NNS IN | 11 | 20 | 2 |
| NN NNS | 10 | 8 | 10 |
| DT NN | 13 | 7 | 7 |
| NN CC | 10 | 11 | 2 |
| NN , | 7 | 14 | 2 |
| IN NNS | 9 | 11 | 2 |

#### Most frequent Trigram-POS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| pos | POI | MP | LP | example |
| NNS IN NN | 4 | 9 | 0 | examples of defect |
| IN DT NN | 4 | 3 | 5 | in the nature |
| IN NN NN | 3 | 9 | 0 | within unit cell |
| NNS IN NNS | 4 | 7 | 0 | impurities on materials |
| NN IN NN | 6 | 5 | 0 | ffect of defect |
| NN CC NN | 5 | 4 | 0 | Stress or deformation |
| NN , NN | 2 | 7 | 0 | toughness , stiffness |
| DT JJ NN | 2 | 4 | 3 | The real world |
| NN NN NN | 2 | 7 | 0 | Unit cell direction |
| NN IN DT | 3 | 2 | 3 | beginning of the |
| NN IN NNS | 4 | 3 | 1 | direction on materials |

Coverage of the Top 10 frequent Trigram-POS

|  |  |  |  |
| --- | --- | --- | --- |
|  | POI | MP | LP |
| top1 | 1.22% | 2.23% | 0.00% |
| top2 | 2.45% | 2.98% | 2.75% |
| top3 | 3.36% | 5.21% | 2.75% |
| top4 | 4.59% | 6.95% | 2.75% |
| top5 | 6.42% | 8.19% | 2.75% |
| top6 | 7.95% | 9.18% | 2.75% |
| top7 | 8.56% | 10.92% | 2.75% |
| top8 | 9.17% | 11.91% | 4.40% |
| top9 | 9.79% | 13.65% | 4.40% |
| top10 | 10.70% | 14.14% | 6.04% |

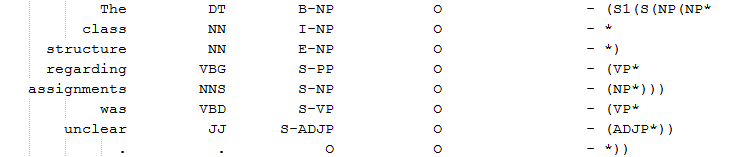
#### Observations:

* According to the distribution of trigram-POS, there is no typical pattern for the phrases. The most common patterns are “NNS IN NN” and “IN NN NN”. However, the coverage for the top 10 frequency trigram patterns are pretty low.

### Phrase-Level

I think the POS is still at a low level so that it has similar sparsity problem as the lexicon. Therefore, I moved to phrase level to see whether there are common pattern among the summaries.

I used the chunk-based phrase extraction. Take the following sentence as an example,



CHUNK

It has 5 phrases: NP (the class structure), PP (regarding), NP(assignments), VP(was), ADJP(unclear)

#### Single Phrase Distribution

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| phrase | POI | MP | LP | example |
| NP | 147 | 170 | 76 | Interactive class |
| PP | 43 | 48 | 23 | on |
| VP | 26 | 36 | 39 | is |
| ADVP | 6 | 13 | 4 | Visually |
| ADJP | 7 | 4 | 7 | good |
| SBAR | 0 | 1 | 2 | if |
| PRT | 0 | 1 | 0 | out |

#### Bigram- Phrase

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| phrase | POI | MP | LP | example |
| PP NP | 41 | 48 | 21 | on graphs |
| NP PP | 36 | 41 | 13 | more explanation on |
| NP NP | 39 | 39 | 5 | more applications real life examples |
| NP VP | 22 | 22 | 25 | Interactive class is |
| VP NP | 16 | 26 | 16 | Need more explanation |
| VP PP | 3 | 5 | 7 | learning from |
| NP ADVP | 5 | 7 | 2 | myself then |
| ADVP NP | 4 | 7 | 0 | Why the strengthening |
| VP ADJP | 4 | 1 | 5 | is good |
| ADVP VP | 1 | 6 | 0 | exactly are |
| NP ADJP | 2 | 3 | 0 | they foremd |
| ADJP PP | 3 | 1 | 1 | very helpful Including |
| PP VP | 2 | 0 | 2 | including explaining |
| ADJP VP | 0 | 2 | 2 | Negative seeing |
| SBAR NP | 0 | 1 | 2 | if they |
| ADVP PP | 1 | 0 | 2 | then within |
| VP VP | 0 | 1 | 1 | turning to see |
| VP ADVP | 0 | 1 | 1 | Reading ahead |
| ADJP NP | 0 | 0 | 1 | Negative Bad Class environment |
| PRT VP | 0 | 1 | 0 | out was |
| ADJP SBAR | 0 | 0 | 1 | helpful if |
| VP SBAR | 0 | 0 | 1 | help if |
| NP PRT | 0 | 1 | 0 | Print out |
| NP SBAR | 0 | 1 | 0 | the graphs if |

#### Trigram-Phrase

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| phrase | POI | MP | LP | example |
| NP PP NP | 34 | 41 | 11 | more explanation on graphs |
| NP VP NP | 13 | 16 | 8 | HW preview problem redoing it |
| NP NP NP | 15 | 17 | 0 | grain size recrystallization properties |
| PP NP NP | 9 | 12 | 2 | from classmates TAs |
| PP NP PP | 9 | 9 | 3 | like the preview problem of |
| VP PP NP | 3 | 5 | 7 | learning from classmates |
| VP NP PP | 4 | 4 | 5 | Need more explanation on |
| NP NP PP | 4 | 7 | 1 | the white board noises from |
| PP NP VP | 5 | 4 | 3 | to Tucson example helped |
| VP NP VP | 4 | 2 | 5 | are explained they are well labled |
| NP VP PP | 2 | 4 | 4 | others learning from |
| NP NP VP | 4 | 5 | 1 | Graphs pictures help |
| NP VP ADJP | 4 | 0 | 5 | Interactive class is good |
| VP NP NP | 2 | 4 | 2 | Seeing more applications real life examples |
| NP ADVP NP | 4 | 4 | 0 | its graphs why ductility |
| ADVP NP VP | 3 | 5 | 0 | Why the strengthening occurs based |
| VP NP ADVP | 1 | 4 | 1 | doing it first |
| ADVP VP NP | 1 | 5 | 0 | exactly are what |

#### Summary Level

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| phrase | POI | MP | LP | example |
| NP | 9 | 14 | 13 | Group activities |
| NP PP NP | 4 | 9 | 1 | Teaching to others |
| NP PP NP PP NP | 4 | 4 | 1 | Don 't like the preview problem of the HW |
| NP VP | 1 | 0 | 4 | Test review helped |
| NP VP NP | 1 | 2 | 2 | Visuals help pictures and graphs |
| NP NP | 4 | 0 | 0 | Cyclotron uranium info |
| NP PP NP NP | 1 | 2 | 0 | Labling phase regions of phase diagram Activity |
| NP PP NP VP NP | 1 | 1 | 0 | Different types of diffusion related processing distinguishing them |
| NP VP NP PP NP | 1 | 0 | 1 | HW preview problem redoing it at the class |
| VP NP | 0 | 2 | 0 | Interpreting the phase diagram graph |
| VP NP PP NP | 0 | 0 | 2 | Need more explanation on graphs |
| NP NP PP NP | 2 | 0 | 0 | The real world examples disasters relation to materials |
| NP PP NP PP NP NP | 2 | 0 | 0 | Differentprocesses of difusion like gear 's case hardening |
| VP PP NP | 0 | 0 | 2 | Listening to teacher |
| NP VP ADJP | 0 | 0 | 2 | Interactive class is good |

NP coverage at summary level: (the ratio of summary is NP compared to all types of summary)

|  |  |  |
| --- | --- | --- |
| POI | MP | LP |
| 18.00% | 23.73% | 30.23% |

#### Observations:

* For this data set, Phrase is definite the right Level
  + 64.2%, 62.3% and 50.3% of the summary is NP, for POI, MP, LP respectively.
  + 57.9% trigram phrases are “NP PP NP”, “NP VP NP”, “NP NP NP”, “PP NP NP”, “PP NP PP”

## Phrase Clustering

One of the key challenge of summarization is to remove duplication.

### K-Means

K-Means is the most popular clustering method and has been used in summarization [1, 2, 3].

Definition:

Given a set of observations (**x**1, **x**2, …, **x***n*), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into *k* (≤ *n*) sets **S** = {*S*1, *S*2, …, *Sk*} so as to minimize the within-cluster sum of squares (WCSS). In other words, its objective is to find:

\underset{\mathbf{S}} {\operatorname{arg\,min}}  \sum_{i=1}^{k} \sum_{\mathbf x_j \in S_i} \left\| \mathbf x_j - \boldsymbol\mu_i \right\|^2 

where ***μ****i* is the mean of points in *Si*.

### Phrase vector

The key for K-Means is the representation of observations. In this case, we need a *d*-dimensional real vector to represent the phrase. I used the word vector. However, instead of using the word itself, I used stemmed word, which shows a better performance in than Soft-NP model than the raw form.

### Clustering Results

I used the python toolkit [Biopython](http://biopython.org/wiki/Main_Page) to implement the K-Means. The distance metric is the default one: Euclidean distance.

The NP extraction is still based on chunk. (I will try syntax one later)

Examples:

For Week1, Muddiest Point

TA’s summary:’

1) Grading process [14]

2) Homework assignments [5]

3) Differences between types of bonding

K=10

|  |  |  |
| --- | --- | --- |
| cluster id | NPs | count |
| 0 | an opportunity, an intro | 2 |
| 1 | a bike " activity, diagram and descriptions, what homework | 3 |
| 2 | a few years, a shape, a bit, a little, a 3d shape, a method | 6 |
| 3 | it, itself, all, need, chem, 't, what, discussion, bonds, those topics, mud, everything, none, today, more, we, no part, that, students, all exercises, extrusion and wire, material, nothing, difference, class, types, chapter, kind, i, bonding, materials, about polymers and bonding, definitions, drawing | 34 |
| 4 | last semester, hw assignments, hw, assignments | 4 |
| 5 | the way, the expectation, the anonymous numbers, the homework, the point, the activity, the use, the sounds, the subject, the class structure, the normalized score, the class, the properties | 13 |
| 6 | molten tin, system, equal fairness, " parts | 4 |
| 7 | the grading system, the grading process, the grading scale, the grading, grading scale | 5 |
| 8 | how normalized grades, grading, grade normalizing | 3 |
| 9 | the test scores, the tests questions, the material, the pre- test, the different material types, the molten tin | 6 |

K=5

|  |  |  |
| --- | --- | --- |
| cluster id | NPs | count |
| 0 | it, 't, an opportunity, last semester, an intro, none, today, more, students | 9 |
| 1 | itself, need, chem, discussion, bonds, hw assignments, equal fairness, we, hw, difference, kind, bonding, " parts, assignments, about polymers and bonding | 15 |
| 2 | a bike " activity, a few years, a shape, a bit, a little, a 3d shape, a method | 7 |
| 3 | molten tin, diagram and descriptions, how normalized grades, all, what, grading, system, those topics, mud, everything, no part, that, all exercises, extrusion and wire, material, nothing, class, types, chapter, grade normalizing, i, grading scale, materials, what homework, definitions, drawing | 26 |
| 4 | the grading system, the test scores, the way, the expectation, the tests questions, the material, the anonymous numbers, the grading process, the pre- test, the homework, the grading scale, the point, the activity, the use, the grading, the sounds, the subject, the different material types, the molten tin, the class structure, the normalized score, the class, the properties | 23 |

#### Observations:

* The problem with clustering algorithm is how to set the K.
  + With K=10, the 3 summaries given by the TA are correctly grouped together, even though they are noisy
  + With K=5, the “assignment” and “bond” merged into one

## Duplication Removing based on clustering

* Rule 1: Select only one phrase from one cluster
* Rule 2: The order of the phrases in one cluster is determined by their Soft-frequency
* ? Rule 3 (not considered): The order of the cluster is determined by the number of phrases in the cluster. (It is not perfect because it doesn’t consider the frequency of the phrases in the cluster.)

## Summarization based Phrase Clustering

### Results

K = 10 is used in the K-Means

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | POI | | | MP | | | LP | | |
|  | R1 | R2 | R-SU4 | R1 | R2 | R-SU4 | R1 | R2 | R-SU4 |
| Unigram | 35.22% | 1.72% | 9.19% | 34.61% | 1.95% | 8.71% | 23.55% | 0.89% | 4.22% |
| TopicS-nostemming | 30.42% | 1.18% | 7.36% | 33.81% | 0.72% | 8.74% | 17.87% | 0.37% | 2.66% |
| TopicS-stemming | 29.98% | 0.92% | 7.04% | 30.91% | 0.27% | 7.16% | 19.14% | 0.37% | 2.88% |
| NP-Hard | 26.98% | 4.30% | 5.85% | 27.75% | 5.77% | 6.44% | 18.11% | 0.16% | 2.67% |
| NP-Soft | 35.83% | 8.74% | 9.95% | 34.11% | 6.92% | 8.74% | 22.43% | 1.93% | 4.00% |
|  |  |  |  |  |  |  |  |  |  |
| SyntaxNP-Hard | 27.55% | 5.43% | 6.65% | 24.67% | 7.71% | 5.85% | 19.35% | 1.13% | 3.48% |
| SyntaxNP-Soft | 37.63% | 9.98% | 11.74% | 32.30% | 10.03% | 8.17% | 22.05% | 2.11% | 3.50% |
|  |  |  |  |  |  |  |  |  |  |
| Cluster-NP-Soft | 34.17% | 10.27% | 10.55% | 33.42% | 9.51% | 9.99% | 18.63% | 1.81% | 4.09% |

### Observations:

* Clustering based method gets higher R2 scores for POI and MP compared to NP-Soft. It also leads to higher R-SU4 scores (the best model for MP).

## TODO

* Syntax-based Cluster NP
* Try other similarity metrics

## Paper I read

Wang, Lu, et al. "A Sentence Compression Based Framework to Query-Focused Multi-Document Summarization." *ACL (1)*. 2013.

Khoury, Richard. "Sentence Clustering Using Parts-of-Speech." *International Journal of Information Engineering and Electronic Business (IJIEEB)* 4.1 (2012): 1.

Check out the annotation with mendeley by clicking the following link.

<http://www.mendeley.com/share/document/invite/7495b1db2a/?utm_medium=email&utm_source=transactional&utm_campaign=share%2Finvitation-document>

Reference:

[1] Lin, Dekang, and Xiaoyun Wu. "Phrase clustering for discriminative learning."*Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*. Association for Computational Linguistics, 2009. [http://nb.mit.edu/f/15283]

[2] Pei-ying, Zhang, and Li Cun-he. "Automatic text summarization based on sentences clustering and extraction." *Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on*. IEEE, 2009.

[3] Wan, Xiaojun, and Jianwu Yang. "Multi-document summarization using cluster-based link analysis." *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008.