# Summary Report (09/18/2014)

## Overview

In this week, I have done the following things:

* Analysis of TA’s summary

## 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% |

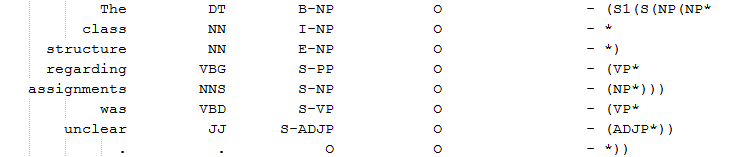
#### Observations:

* 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

#### Bigram- Phrase and Trigram-Phrase

## Phrase Clustering

### K-Means

### Distance Metrics

### Clustering Results

## Duplication Removing based on clustering

## Summarization based Phrase Clustering

### Approach

### Results

## Paper I read

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

[2] 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>