# Summary Report (07/31/2014)

## Overview

In this week, I have done the following things:

* Implemented several baselines and compare them with Mead
* Implemented several shallow summarizer and compare them with Mead

## Baselines

All the following results are the preprocessing:

* Responses splitting: a student’s response will be split if it has several sentences with the NLTK sentence tokenizer
* Response filtering: the following responses are filtered out: "?", "[blank]", 'n/a', 'blank' (I didn’t filter out the responses like “nothing”, “none” because I think these responses are also good, which could be detected with a classifier to predict whether the student has a point or none)

In addition, I added the Rouge-1 (R1) evaluation metric because the bigram coverage (shown in last report) is not very good but the unigram coverage is good.

### Random baseline

For each of summarization, the model outputs K random sentences from the students’ responses (each sentence has an equal probability to be selected). K is an integer to represent the sentence compression rate. In the experiments, K ranges from 1 to 3.

### Longest sentences baseline

In this model, instead of picking up K random sentences, it outputs the K sentences which have the most number of words.

### Shortest sentences baseline

Another straightforward baseline is to always pick up the shortest K sentences.

### Results

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | POI | | | MD | | | LP | | |
|  | R1 | R2 | R-SU4 | R1 | R2 | R-SU4 | R1 | R2 | R-SU4 |
| default mead | 25.07% | 8.33% | 5.43% | 25.03% | 8.51% | 6.23% | 14.47% | 1.72% | 2.28% |
| Random(K=3) | 22.80% | 5.21% | 4.98% | 16.47% | 3.50% | 2.75% | 15.59% | 3.73% | 2.68% |
| Random(K=2) | 14.31% | 1.72% | 2.27% | 20.04% | 4.59% | 3.02% | 12.47% | 2.66% | 2.17% |
| Random(K=1) | 14.37% | 2.44% | 2.68% | 11.02% | 3.11% | 1.55% | 6.64% | 0.00% | 0.63% |
| Longest(K=3) | 19.74% | 4.39% | 4.09% | 23.86% | 6.92% | 4.75% | 13.12% | 0.61% | 1.48% |
| Longest(K=2) | 19.27% | 4.34% | 3.64% | 23.24% | 6.35% | 4.43% | 12.17% | 0.68% | 1.40% |
| Longest(K=1) | 16.01% | 3.65% | 2.74% | 18.58% | 4.50% | 3.31% | 10.09% | 0.64% | 1.62% |
| Shortest(K=3) | 12.11% | 0.57% | 1.81% | 7.49% | 0.85% | 0.98% | 7.87% | 0.00% | 1.03% |
| Shortest(K=2) | 8.68% | 0.00% | 1.00% | 5.96% | 0.46% | 0.49% | 3.34% | 0.00% | 0.11% |
| Shortest(K=1) | 4.45% | 0.00% | 0.00% | 2.89% | 0.00% | 0.03% | 1.19% | 0.00% | 0.00% |

\* Highlighted cells are significantly different from the default mead (p<0.05, using pair ttest)

### Observation:

* It turns out the default mead is not significantly different from the random baseline with K=3 except MP. It means that the mead is not very good for POI and LP.
  + After a close look at the numbers (the details for the results for the random model and mead model are shown below: red numbers are bad ones), the major issue is that the R2, R-SU4 for many weeks in the mead summary are very low. Thus, even if the mead summary is much better for some of the weeks than the random baseline (such as week 3, 9, 11), it is not significantly better overall.
* The Longest and Shortest baselines are bad, especially when K = [1, 2]. It makes sense that the shortest baseline is bad because the recall is too low. The longest baseline is also bad because with longer sentence, the precision become worse. (Rouge score is the f-measure.)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mead | POI | | | MP | | | LP | | |
| week | R1 | R2 | R-SU4 | R1 | R2 | R-SU4 | R1 | R2 | R-SU4 |
| 1 | 16.67% | 0.00% | 1.69% | 3.08% | 0.00% | 0.12% | 4.55% | 0.00% | 0.32% |
| 2 | 19.05% | 9.84% | 2.63% | 29.36% | 13.08% | 5.04% | 11.94% | 0.00% | 0.78% |
| 3 | 26.47% | 21.21% | 6.93% | 24.00% | 10.96% | 3.14% | 12.99% | 0.00% | 1.44% |
| 4 | 16.67% | 0.00% | 2.23% | 30.56% | 8.57% | 8.03% | 3.77% | 0.00% | 0.22% |
| 5 | 22.95% | 10.17% | 4.79% | 28.57% | 6.25% | 9.65% | 12.66% | 0.00% | 1.13% |
| 6 | 21.28% | 0.00% | 3.92% | 27.27% | 4.76% | 6.34% | 22.58% | 6.67% | 3.37% |
| 7 | 25.97% | 5.33% | 5.90% | 22.54% | 14.49% | 5.41% | 6.25% | 0.00% | 0.30% |
| 8 | 27.45% | 8.16% | 6.18% | 20.69% | 7.06% | 3.08% | 16.00% | 0.00% | 1.52% |
| 9 | 27.37% | 6.45% | 6.22% | 40.34% | 10.26% | 14.16% | 23.53% | 4.08% | 3.60% |
| 10 | 26.42% | 3.92% | 7.31% | 30.00% | 12.82% | 8.30% | 30.77% | 8.00% | 9.69% |
| 11 | 45.76% | 32.76% | 13.84% | 15.04% | 1.53% | 1.66% | 21.62% | 1.84% | 4.62% |
| 12 | 24.74% | 2.11% | 3.56% | 28.92% | 12.35% | 9.79% | 7.02% | 0.00% | 0.35% |
| average | 25.07% | 8.33% | 5.43% | 25.03% | 8.51% | 6.23% | 14.47% | 1.72% | 2.28% |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Random | POI | | | MP | | | LP | | |
| week | R1 | R2 | R-SU4 | R1 | R2 | R-SU4 | R1 | R2 | R-SU4 |
| 1 | 25.64% | 0.00% | 5.26% | 24.00% | 0.00% | 4.47% | 22.73% | 0.00% | 3.52% |
| 2 | 9.76% | 0.00% | 0.85% | 25.40% | 6.56% | 5.08% | 24.24% | 0.00% | 5.01% |
| 3 | 15.79% | 5.56% | 2.42% | 17.39% | 4.55% | 1.77% | 20.90% | 0.00% | 4.50% |
| 4 | 24.24% | 0.00% | 4.01% | 33.33% | 15.00% | 7.86% | 35.90% | 10.81% | 8.18% |
| 5 | 4.44% | 0.00% | 0.35% | 10.35% | 3.57% | 1.21% | 12.77% | 4.44% | 2.07% |
| 6 | 32.00% | 0.00% | 4.73% | 42.86% | 7.69% | 19.23% | 18.61% | 0.00% | 3.19% |
| 7 | 29.85% | 6.15% | 9.00% | 20.83% | 0.00% | 3.67% | 5.41% | 0.00% | 0.41% |
| 8 | 36.36% | 25.81% | 7.30% | 27.91% | 19.51% | 6.64% | 27.59% | 7.41% | 8.00% |
| 9 | 12.50% | 0.00% | 1.13% | 9.20% | 4.71% | 0.67% | 5.41% | 0.00% | 0.45% |
| 10 | 21.05% | 11.11% | 4.01% | 26.47% | 6.06% | 4.06% | 26.09% | 4.55% | 6.96% |
| 11 | 20.20% | 6.19% | 2.71% | 27.69% | 7.81% | 5.44% | 7.84% | 0.00% | 0.60% |
| 12 | 33.33% | 17.14% | 5.25% | 16.33% | 0.00% | 2.81% | 5.88% | 0.00% | 0.23% |
| average | 22.10% | 6.00% | 3.92% | 23.48% | 6.29% | 5.24% | 17.78% | 2.27% | 3.59% |

## Shallow Summarization

Task definition: select the top N keywords (N=30, because the average number of words in the students’ response and TA’s summary are both close to 10 and I assume 3 points will be extracted); If it is to select key phrases, the total number of words should be less or equal to N.

### N-gram frequency

* Unigram model: rank the unigrams by their frequency
* Ngram model: rank all the ngrams (up to 5-gram)
* Weighted Ngram model: In the previous ngram model, each appearance of a ngram counts only 1. However, longer ngram should be more valuable. For example, “main topic of the class” is more valuable than “class”. Thus, each ngram gives the weight as the number of words. In this case, “main topic of the class” has weight 5. In this case, this model should prefer longer ngram because if longer ngram occurs multiple times in the student’s responses, they will have more weights. For example, if “main topic of the class” appears twice, it will have a weight of 10.
* Bigram model: To favor the R2 metric, only bigram is considered.
* Bigram with K=60: for the previous bigram model, it will produce a list of 30 bigrams as the summary. To test whether increase the length of the output list will also increase the R2 score, this model changed the K to 60.

#### Results:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | POI | | | MP | | | LP | | |
|  | 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 | 35.01% | 1.97% | 9.60% | 36.59% | 2.41% | 9.93% | 24.48% | 0.39% | 4.65% |
| ngram | 32.61% | 3.18% | 9.49% | 36.16% | 6.66% | 11.22% | 21.72% | 1.29% | 3.79% |
| weighted | 25.47% | 5.20% | 6.75% | 31.48% | 10.33% | 9.54% | 12.10% | 1.72% | 1.96% |
| bigram | 29.25% | 7.97% | 7.21% | 36.43% | 12.74% | 10.60% | 14.75% | 2.08% | 2.40% |
| bigram(K=60) | 28.42% | 7.74% | 6.93% | 32.32% | 10.33% | 8.84% | 14.55% | 2.20% | 1.91% |

#### Observations

* This new type of summary framework changed the ROUGE scores significantly
  + The unigram gets significantly better R1 and R-SU4 scores on all the POI, MP and LP
  + However, it gets lower score for R2 (which makes sense)
* The Unigram model has best R1 score because with limited output word number, the unigram model can maximize the coverage of the unigrams and probably many of the unigrams in the TA’s summary also belong to the most frequent ones
* The Ngram model get significantly better R2 score than the Unigram model. However, the R2 score is still lower than the Mead.
* The Ngram model’s R-SU4 are better, which is close to the level of the state-of-the-art R-SU4 score (except LP).
* Weighted model doesn’t work as expected. (Some of the insights are shown in the example below)
* Increasing the output summary length for the bigram model doesn’t yield better R2 score, probably it produces more noisy outcome.

#### Ngram Examples

TA’s summary:

For Week 3, POI

* How bond strength is related to melting point [10]
* Relative strenghts of different bond types [8]
* Elastic modulus [5]
* How plastic bags are made by polymer fil blowing [4]

For Week 1, LP

* Activities [10]
* Reading ahead [4]
* Negative: Bad Class environment (like turning to see the board, not being able to read the white board, noises from other classes, etc…) [4]
* Graphs are helpful if they come along side text [3]

Ngram frequency (yellow ones are good; red ones are bad)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Good (week3, POI) | |  | Bad (week1, LP) | |
| ngram | frequency |  | ngram | frequency |
| the | 25 |  | i | 23 |
| of | 22 |  | to | 17 |
| bond | 13 |  | learn | 12 |
| i | 12 |  | the | 12 |
| to | 9 |  | and | 8 |
| how | 8 |  | in | 8 |
| melting | 8 |  | learned | 7 |
| on | 8 |  | group | 7 |
| , | 7 |  | i learn | 7 |
| properties | 7 |  | learning | 7 |
| strength | 6 |  | i learned | 6 |
| that | 6 |  | groups | 5 |
| and | 6 |  | my | 5 |
| bonding | 6 |  | like | 5 |
| interesting | 6 |  | well | 5 |
| is | 5 |  | of | 5 |
| between | 5 |  | i like | 4 |
| learning | 5 |  | a | 4 |
| different | 5 |  | better | 4 |
| a | 5 |  | that | 4 |
| are | 5 |  | class | 4 |
| point | 5 |  | in groups | 4 |
| type | 4 |  | - | 4 |
| in | 4 |  | by | 4 |
| . | 4 |  | working | 4 |
| it | 4 |  | other | 4 |
| modulus | 4 |  | activities | 3 |
| was | 4 |  | me | 3 |
| thermal expansion | 4 |  | notes | 3 |
| learning about | 4 |  | what | 3 |
| bags | 4 |  | , | 3 |
| type of | 4 |  | learned that | 3 |
| melting point | 4 |  | not | 3 |
| thermal | 4 |  | i learned that | 3 |
| temperature | 4 |  | doing | 3 |
| expansion | 4 |  | ' | 3 |
| about | 4 |  | way | 3 |
| made | 4 |  | to learn | 3 |
| bond strength | 4 |  | learn better | 3 |
| + | 3 |  | best | 3 |
| types | 3 |  | be | 3 |
| type of bond | 3 |  | on | 3 |
| of bonding | 3 |  | with | 3 |
| materials | 3 |  | as | 3 |
| were | 3 |  | for | 3 |
| strengths | 3 |  | all | 2 |
| elastic modulus | 3 |  | activities as | 2 |
| polymer | 3 |  | helps | 2 |
| i found | 3 |  | small | 2 |

#### Observations:

* This shallow summary frame work has two limitations
  + First, the output are phrases rather than sentences. And the number of output units are much larger than traditional extractive summary
  + Second, it considers only the word’s surface form, but doesn’t consider the relations between them. (duplicate, representative)
    - In the above example, “learn”, “I learned”, “I learned that”, “learned that”, all of them appears in the list with high ranks. But they should be consider as only one [may be clustering].
    - Some of the ngrams have no meaning, like the punctuations, prepositions “in”, “to” [should be treated as stop words and filter out]

### Ngrams with removing the stopwords

In these models, they are nearly the same as the previous ones, however, now, ngrams that start with a stop word (also the punctuations) will not be considered.

The stop word list is from the ROUGE toolkit.

The punctuation list is ['.', '?', '-', ',', '[', ']', '-', ';', '\'', '"', '+', '&', '!', '/', '>', '<', ')', '(', '#', '=']

#### Results

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | POI | | | MP | | | LP | | |
|  | 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% |
| ngram | 30.62% | 5.21% | 8.88% | 35.02% | 10.51% | 10.14% | 21.44% | 1.25% | 4.38% |
| weighted | 25.63% | 6.30% | 6.17% | 29.32% | 10.93% | 8.28% | 16.62% | 2.53% | 3.18% |
| bigram | 31.89% | 8.16% | 8.99% | 32.34% | 11.01% | 8.47% | 15.97% | 1.89% | 2.80% |

#### Observations

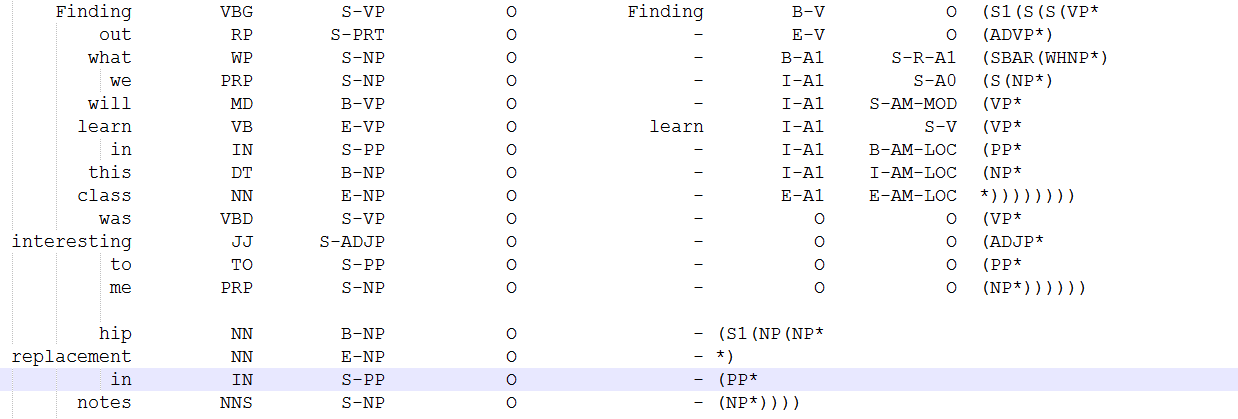
* Overall speaking, there are not very different from the old models
  + It just solved part of the second limitation. The output list still has overlaps.
* Here are some minor changes
  + It doesn’t have a significant loss for the R1
  + The bigram model improves, especially for the R-SU4 for POI
  + The R2 of ngram model are no longer significantly different from the mead now.

### NP Phrase frequency

Even though with removing the stop words, the output list still have phrases like “like that”, “found that”… which is not useful. Thus, the idea of the NP phrase model is to just consider the noun phrases.

Now, this model is implementing. Now, I successfully get the chunk parser with the [Senna toolkit](http://ml.nec-labs.com/senna/).

Here is an example of the output:



The next step is to extract the noun phrases.

## Talks I attended

### Zahra’s Practice talk Preliminary exam

### Randomized Controlled Trail Using ASSISTments

Interesting points

* It is a web-based talk based on <https://connect.wpi.edu>.
  + It supports camera, microphone, shared slides
    - The meeting organizers have the right to let who can speak
    - Other attendance can join the type-based conversation
* ASSISTment
  + It is a web-based user study platform
    - The most interesting for this tool is that it has built-in support for random-control, multiple type of resources (video, picture) by a word-like editor.
    - In addition, it can serve as a simple tutoring system (instructors can give correct answer (also the hints) to the assignments so that the students can receive immediate feedback)

## Paper I read

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>]