# Summary Report (08/07/2014)

## Overview

In this week, I have done the following things:

* Topic Signature Summary: both shallow and extractive
* NP-Phrase based Summary: shallow

## Topic Signature Summary

### The idea of topic words for summarization was introduced by (Lin and Hovy 2000). A topic word is a word with significantly greater probability in a given text compared to that in a large background corpus.

Here I used the software TopicS implemented by Annie Louis (<http://homepages.inf.ed.ac.uk/alouis/topicS.html>) to extract the topic words (The same software used by Wenting).

The input of the TopicS software are the documents to be extracted and a set of background documents (5000 documents from the English GigaWord Corpus); the output is a list of topic words, associated with the chisq values (the bigger,, the more important.)

### Shallow Summary

In this setting, topic signature is used as a way to select keywords. Like the experiment I did using the unigram, the output is a list of words/phrases and the total number of words is not bigger than K.

#### Results

I tested the TopicSignature under two conditions: with stemming and without stemming

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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% |

#### Examples of Topic Signatures vs. unigram:

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]

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| unigrams | | | |  | topic signatures | | | |
| (class3, POI) | | (class 1, LP) | |  | (class 3, POI) | | (class 1, LP) | |
| unigram | frequency | unigram | frequency |  | topic words | Chi-sq | topic words | Chi-sq |
| bond | 13 | learn | 12 |  | melting | 52.0578 | learn | 52.9179 |
| melting | 8 | group | 7 |  | bond | 50.6245 | learning | 33.2103 |
| properties | 7 | learned | 7 |  | bonding | 34.2792 | learned | 27.7005 |
| interesting | 6 | learning | 7 |  | properties | 33.1987 | group | 16.8917 |
| strength | 6 | groups | 5 |  | made | 30.8132 | groups | 14.954 |
| bonding | 6 | working | 4 |  | modulus | 30.8132 | class | 11.6441 |
| point | 5 | class | 4 |  | interesting | 24.2747 | activities | 11.3693 |
| learning | 5 | activities | 3 |  | strength | 23.1578 | interaction | 10.799 |
| thermal | 4 | notes | 3 |  | found | 23.104 | notes | 10.0823 |
| expansion | 4 | knowledge | 2 |  | polymer | 23.104 | working | 10.0285 |
| temperature | 4 | helps | 2 |  | thermal | 22.7585 | vulcanization | 8.0142 |
| type | 4 | material | 2 |  | learning | 20.6752 | discussions | 7.7904 |
| made | 4 | small | 2 |  | bags | 17.6072 | kinesthetic | 6.8101 |
| bags | 4 | enjoy | 2 |  | elastic | 17.5065 | deduce | 6.8101 |
| modulus | 4 | people | 2 |  | temperature | 15.7138 | knowledge | 6.8092 |
| found | 3 | discussions | 2 |  | ionic | 15.3987 | helps | 6.7977 |
| elastic | 3 | pictures | 2 |  | covalent | 15.3987 | enjoy | 6.7306 |
| polymer | 3 | interaction | 2 |  | waals | 15.3987 | teamwork | 6.3557 |
| materials | 3 | makes | 2 |  | expansion | 15.2861 | pictures | 6.1734 |
| strengths | 3 | work | 2 |  | strengths | 14.5466 | visuals | 5.8413 |
| types | 3 | taking | 2 |  | coefficient | 13.7403 | interacting | 5.8413 |
| covalent | 2 | things | 2 |  | polymers | 12.9907 | complicating | 5.5211 |
| microscopic | 2 | rounded | 1 |  | atoms | 12.9907 | repetition | 5.1913 |
| bonds | 2 | obtain | 1 |  | type | 12.7115 | enhances | 5.0254 |
| polymers | 2 | repetition | 1 |  | pt | 12.082 | lectures | 4.7121 |
| der | 2 | personal | 1 |  | materials | 10.8896 | rounded | 4.5672 |
| energy | 2 | teaching | 1 |  | microscopic | 10.8697 | chemistry | 4.4826 |
| waals | 2 | resources | 1 |  | types | 10.7842 | lecture | 4.4055 |
| learned | 2 | answer | 1 |  | der | 10.5523 | taking | 4.3382 |
| coefficient | 2 | hands | 1 |  | point | 10.4204 | assignments | 4.2383 |

#### Observations:

* Shallow summary based on topic words doesn’t show better results in terms of ROUGE scores
* The overlap between frequency-ranked unigram and topic signature is large
  + Probably the default background document is not good fit for this task (the survey is more like oral conversation. It’s informal. Thus, the topic signature cannot filter out the word like “found”)
* The topicSignature is still based one individual word, therefore, it is bad on the R2 score.

### Extractive Summary

A summary with a list of words is not friendly to users because it is not easy for the TA/instructor to understand the words. Thus, we proposed the word-based extractive summary.

The basic idea is simple. Since we already have the list of words and their weights (either using the unigram and frequency or using the topic signature words and chi-sq scores), we can select the sentence that has the most coverage score.

The sentences are selected greedily as the follows:

1. Given a document D, Compute the score function f(w) for each word
2. Initialize the summary sentences result set as R={}, initial word set
3. Select the sentence S that has the largest score. R’ = R + {S}
4. Delete the words in S from the word set,
5. Go back to step 2, until a certain number of sentences are selected

Note: the step 4 is aim to reduce the duplication. In an addition, this greedy algorithm could be extended to ILP (Integer linear programming) (the state-of-the-art) to get a global solution with more time.

Weakness: might prefer longer sentence, especially for the first one

#### Results

Parameter setting:

* All of words are stemmed (porter stemming).
* In the unigram based summary, the unigrams that appeared only once are not considered.
* In the topic signature based summary, the words with chi-sq score less than 5 are not considered.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | POI | | | MP | | | 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% |
| unigram-based | 26.71% | 9.02% | 5.57% | 31.39% | 11.60% | 7.97% | 15.46% | 1.35% | 2.18% |
| topicS-based | 28.53% | 9.70% | 7.18% | 31.80% | 9.78% | 8.70% | 16.04% | 1.56% | 2.27% |

#### Observations:

* Both the method outperform the mead in general, especially R1 and R2 for MP with unigram-based method, and R1 and R-SU4 for MP with topicS-based method, which are significantly different from the mead.
* The p-value for R1 and R-SU4 for POI with topicS are close to 0.05 too: 0.061 and 0.055
* The TopicS seems not work (probably due to the selection of background documents) [Another good background document?]
  + The setting for this task is different from others. For typical summarization, the content that is different from the background document is informative. However, in our task, if we choose the material in the lecture/book (e.g. the slides from the instructor), the content that is related to background documents are useful.

#### Examples

|  |  |  |
| --- | --- | --- |
|  | week3 POI | Week 1 LP |
| TA | How bond strength is related to melting point Relative strenghts of different bond typesElastic modulus How plastic bags are made by polymer fil blowing | Working in a small group Picures and visuals Teaching to others |
| Mead | Learning about what bond correspond to what materials  I like learning about the microscopic explanation for the microscopic properties, especially, I liked learning about bond strengths  How bond strength is related to melting point and that melting point is on the periodic table | I learned that taking risks in learning help me learn more I learn best by taking my own notes I learned that I learn best by seeing things presented on the power point |
| Unigram-based | I thought it was interesting that the type of bond makes such a big difference in the temperature required to process different materials How bond strength is related to melting point and that melting point is on the periodic table I like learning about the microscopic explanation for the microscopic properties, especially, I liked learning about bond strengths | I learned that taking risks in learning help me learn more I like group discussions & learning, but not group assignments I enjoy working with other people and doing hands on activities |
| TopicS-based | How bond strength is related to melting point and that melting point is on the periodic table I found the polymer film blowing interesting I had no clue on how bags were made I like learning about the microscopic explanation for the microscopic properties, especially, I liked learning about bond strengths | I learned that taking risks in learning help me learn more I like group discussions & learning, but not group assignments I enjoy working with other people and doing hands on activities |

## NP-Phrase Summary

There are at least two issues for word-based summarization.

* It is ambiguous. For example, the word “point” can refer to the “melting point” and “grading point”. Consider phrase can reduce such ambiguity.
* It might select content that are common in the document but actually they are not informative, such as “I learned”, “I found”, “I like”, etc.

Therefore, we proposed “phrase-based” summarization, in particular, NP phrases are extracted using the chunk parser provided in Senna.

### Phrase-based Shallow Summary

Although phrases are less ambiguous than words, it suffers from sparsity problem. Thus, the frequency cannot be directly computed because many of them appear only once.

We call the straightforward frequency based method as “hard frequency model” (**NP-Hard**). To address this issue, a relax model is proposed (**NP-Soft**). The basic idea is to count the NP phrases without extract matching, instead, if two phrases are overlap in term of stemmed word, and the weight of them are the overlap count between them. For example, Overlap(“interactive groups”, “group interaction”) = 2, Overlap(“pictures and visuals”, “pictures and teamwork”) = 1. The time complexity to construct such a dictionary is O(n^2), n is the number of phrases.

With the relax frequency count model, the similar phrases will cluster into together with similar weights.

For example, the weight dictionary for LP of week1 is shown below.

|  |  |
| --- | --- |
| NP | relax frequency |
| groups | 32 |
| i | 23 |
| group discussions & learning | 17 |
| interactive groups | 13 |
| group interaction | 13 |
| small group communication | 12 |
| small group work | 11 |
| , but not group assignments | 11 |
| a group | 10 |
| group | 10 |
| this class | 7 |
| learning | 4 |
| activities | 4 |
| my learning | 4 |
| kinesthetic learning | 4 |
| a more well- rounded learning experience | 4 |
| activity | 4 |
| any confusing / complicating material | 3 |
| my best way | 3 |
| what | 3 |
| different materials | 3 |
| my own notes | 3 |
| chemistry class | 3 |
| me | 3 |
| the class | 3 |
| it | 2 |

#### Duplicate Removing

However, another problem is the duplication. The question, how to select the phrase. This could be done with [maximum coverage problem](http://en.wikipedia.org/wiki/Maximum_coverage_problem). Current, I only implement a greedy algorithm: selection the phrases from the top to down and ensure that no two phrases are overlap.

#### Results

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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% |

#### Observations:

* NP-Soft improve the R2 a lot compared to the Unigram baseline
* The NP-Soft doesn’t hurt the R1 and get a light improvement over R-SU4.

### Phrase-based Extractive Summary

TODO

## Paper I read

[1] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. 2004. [<http://nb.mit.edu/f/15325>]

[2] Nenkova, Ani, Sameer Maskey, and Yang Liu. "Automatic summarization."*Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts of ACL 2011*. Association for Computational Linguistics, 2011. [<http://nb.mit.edu/f/15326>]

# Reference

C. Lin and E. Hovy. 2000. The automated acquisition of topic signatures for text summarization. In roceedings of the 18th conference on Computational linguistics.