# Summary Report (09/03/2014)

## Overview

In this week, I have done the following things:

* Syntax-based NP Extracting

## Syntax-based NP Extracting

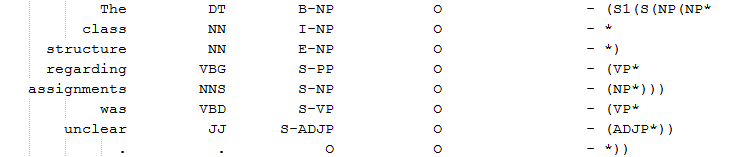
Chunk-based NP extracting is still at a small granularity, because a large NP will be split into several parts.

For example, “the class structure regarding assignments was unclear” has two NP based on chunk:

“the class structure” and “assignments”.

However, when using syntax parser tree, it will merge to one bigger NP

“the class structure regarding assignments”.

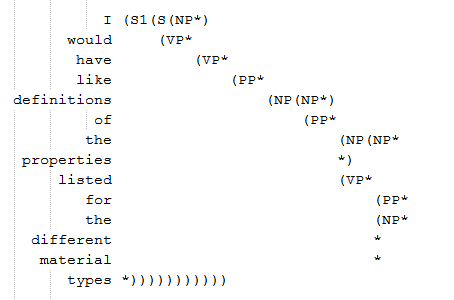


CHUNK Syntax

### NP Extracting

The syntax grammar is represented as a tree marked by parenthesis.

The issue to extract the NP is that they are nested with each other. Take the following sentence as example:  
“I would have like definitions of the properties listed for the different material types” with the following syntax parse tree:



There are 6 NPs in the sentences (highlighted above): (appears in order)

* I
* Definitions of the properties listed for the different material types
* Definitions
* the properties listed for the different material types
* the properties
* the different material types

For this issue, I just take the biggest NP (the top most level) if they are nested. For the case above, the NP “I” and “Definitions of the properties listed for the different material types” will be extracted.

I used a stack to implement the algorithm.

### Examples

For Week1, Muddiest Point

TA’s summary:’

1) Grading process [14]

2) Homework assignments [5]

3) Differences between types of bonding

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SyntaxNP-Hard | | SyntaxNP-Soft | | ChunkNP-Hard | | ChunkNP-Soft | |
| i | 13 | grading | 21 | i | 13 | grading | 19 |
| it | 5 | grading scale | 21 | it | 8 | grading scale | 19 |
| the grading | 3 | a little confused about the grading system still | 18 | nothing | 5 | the grading | 16 |
| a little | 2 | the grading | 17 | a little | 4 | i | 13 |
| grading | 2 | the grading scale | 13 | what | 3 | the grading system | 12 |
| nothing | 2 | i | 13 | bonds | 3 | the grading scale | 12 |
| grading scale | 2 | normalized grades | 12 | the grading | 3 | how normalized grades | 11 |
| everything | 2 | the way we are graded in class | 12 | the pre- test | 2 | grade normalizing | 11 |
| students equal fairness | 1 | grade normalizing | 12 | grading | 2 | the grading process | 9 |
| the test scores | 1 | the grading process | 10 | everything | 2 | it | 8 |
| the sounds of it | 1 | material that was given on the pre- test | 8 | today | 2 | bonds | 6 |
| a little vague | 1 | the pre- test on bonds | 8 | we | 2 | the different material types | 6 |
| a few years | 1 | the material about polymers and bonding . | 7 | that | 2 | the pre- test | 5 |
| the expectation | 1 | nothing about materials | 7 | grading scale | 2 | nothing | 5 |
| 't | 1 | an intro now , so nothing was confusing for today | 7 | a bike " activity | 1 | the tests questions | 4 |
| the material about polymers and bonding . | 1 | nothing | 7 | the grading system | 1 | a little | 4 |
| the grading process | 1 | definitions of the properties listed for the different material types | 6 | the test scores | 1 | the normalized score | 4 |
| the class structure regarding assignments | 1 | a little | 5 | molten tin | 1 | the test scores | 3 |
| chem | 1 | it | 5 | diagram and descriptions | 1 | molten tin | 3 |
| a little confused about the grading system still | 1 | types of bonds | 5 | a few years | 1 | the material | 3 |

Note:

The raw count for single word are not the same as the count in Chunk Based NP, such as “it”. It is because not all the single word are identified as “NP” due to the nesting issue.

### 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% |
|  |  |  |  |  |  |  |  |  |  |
| 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% |

Observation:

* The Syntax based NP model improves the R2 scores a lot. It is even better than the chunk based NP-Soft model.

## 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 Frequency Distribution
* POS Distribution
* Phrase Type

### Word Frequency

### POS

### Phrase

## TODO

A poster for “Wine and Research” event.

## Paper I read

[1] Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. "Building a large annotated corpus of English: The Penn Treebank." *Computational linguistics* 19.2 (1993): 313-330. [http://nb.mit.edu/f/16062]