

Data Preprocessing – Task 5

Visit the Dome

Ruta Bakhda, Thang Nguyen, Vamsi Vaddi

Tasks

Task 1 : To set up annotation tool WAT-SL

Task 2 : Segment the speech into sentences

Task 3 : To set up fragmented sentences in Task 2 in WAT-SL

Task 1 : To set up WAT-SL

WAT-SL :

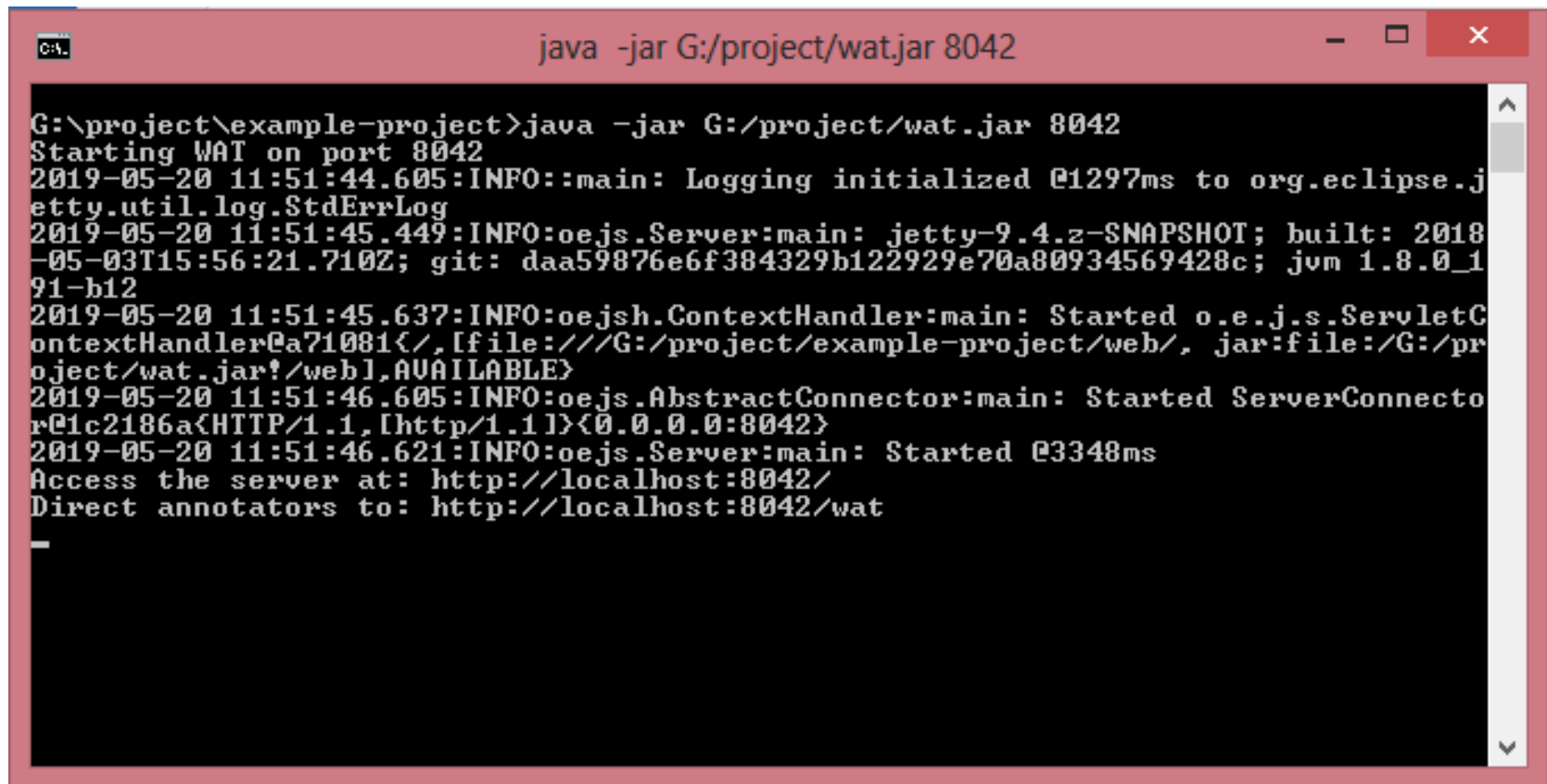
- ▶ Web Annotation tool for Segment Labeling
- ▶ Developed by Webis group
- ▶ Available open source at github.com/webis-de/wat. Default can be configured using wat.jar

▶ *Main 3 interface in WAT-SL*

- ▶ Annotation interface
- ▶ Admin Interface
- ▶ Curation Interface

First step

- ▶ Download wat.jar file and run as below
- ▶ `java -jar <path-to>/wat.jar [<port> [<base-path>]]`

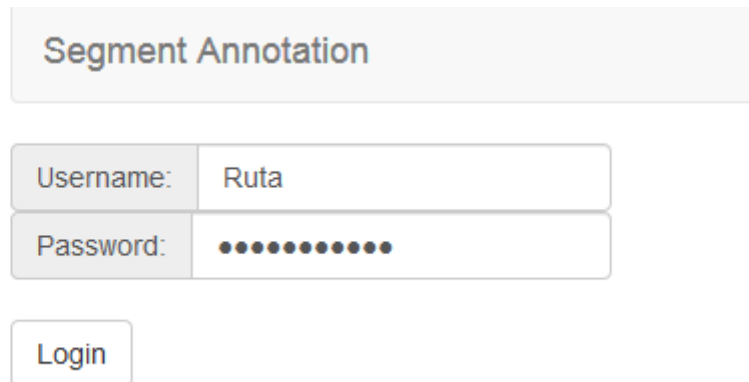


```
java -jar G:/project/wat.jar 8042

G:\project\example-project>java -jar G:/project/wat.jar 8042
Starting WAT on port 8042
2019-05-20 11:51:44.605:INFO::main: Logging initialized @1297ms to org.eclipse.jetty.util.log.StdErrLog
2019-05-20 11:51:45.449:INFO:oejs.Server:main: jetty-9.4.z-SNAPSHOT; built: 2018-05-03T15:56:21.710Z; git: daa59876e6f384329b122929e70a80934569428c; jvm 1.8.0_191-b12
2019-05-20 11:51:45.637:INFO:oejsh.ContextHandler:main: Started o.e.j.s.ServletContextHandler@a71081</, [file:///G:/project/example-project/web/, jar:file:/G:/project/wat.jar!/web/,AVAILABLE>
2019-05-20 11:51:46.605:INFO:oejs.AbstractConnector:main: Started ServerConnector@1c2186a<HTTP/1.1, [http/1.1]><0.0.0.0:8042>
2019-05-20 11:51:46.621:INFO:oejs.Server:main: Started @3348ms
Access the server at: http://localhost:8042/
Direct annotators to: http://localhost:8042/wat
```

1. Annotation Interface

- ▶ Can be accessed at *wat/annotate*
- ▶ Annotators can login using given username and password
- ▶ Shows the list of available task with user's progress on them
- ▶ All progress is saved automatically
- ▶ User configuration can be updated in *wat.conf*



The screenshot shows a web interface for 'Segment Annotation'. It features a login form with two input fields: 'Username' containing the text 'Ruta' and 'Password' containing ten dots. Below the password field is a 'Login' button.

Segment Annotation	
Username:	Ruta
Password:	••••••••••
<button>Login</button>	

User's annotation interface with task list

Segment Annotation Ruta

logout

Choose a task below in order to work on it. All progress is saved automatically. You can return to this page at any time.

You have completed 0 of 4 tasks.

Task	Progress	Last update
002-actorartistphilanthropistmymot	0/64 segments	-
003-greecesuffereuroexitsinglecurr	0/141 segments	-
005-treatingafricanswithanunte	0/111 segments	-
sample	1/119 segments	2019-05-18 14:37



User's annotation interface

All progress is saved automatically. You can return to the task selection at any time.

Task details

Actor, artist, philanthropist: My mother's selfless love with me at the holidays and always. (64 segments remaining)

The holidays are my favorite time of year. As my home fills with the joy and laughter of my six children and four grandchildren, I'm reminded of how lucky I am to have the most wonderful job in the world - being a mother and grandmother. It's the single most important thing I've done with my life, and I know just how important the job is because I am constantly reminded of what an influence my own mother was in my life. My mother Mieke survived three and a half horrific years in a Japanese concentration camp during World War II. During that time, she focused all of her energy on caring for fellow prisoners who were far worse off than she was.

When I think of my mother, I think of the selfless love she had for complete strangers as well. She's with me every day as I challenge myself to live my life in the same way. I am reminded of her each time a single white feather floats through the air - whether on good days or bad days reflected in my work as an actor, artist and philanthropist. Without her example, I know that if I live my life the way she did, that is truly a gift.

CO: Common Ground

AS: Assumption

TE: Testimony

ST: Statistics

AN: Anecdote

OT: Other

No unit

Continued

Labels for annotation

2. Admin Interface

- ▶ Reload Server
 - ▶ Updates all configurations and reads all tasks again

- ▶ Write Results
 - ▶ For all completed tasks, it can be mapped from server directory to 'results' directory in the system

- ▶ Show annotation progress
 - ▶ Allows admin to see progress of each user and each task manually

Segment Annotation

Admin

Password:

●●●●●●●●●●

Login

Segment Annotation

logout

Reload server

Updates all configurations and reads again all tasks

Write results

Writes the 'key = value' mappings of all completed tasks to the 'results' directory in the server directory.

Show annotation progress

logout

Annotation Progress

Annotator	Tasks		Time	Login (new window)
Alice Carrol (alice)	1/3	33%	0:05:29	Login as Alice Carrol
Charlie Brown (charlie)	1/4	25%	0:05:34	Login as Charlie Brown
Ruta (ruta)	0/4	0%	0:00:53	Login as Ruta

When logged in using the buttons above, record files will show 'ADMIN' instead of 'ANNOTATOR' and all actions performed will be ignored when calculating the time spent.

How to configure WAT-SL?

- ▶ Wat.conf (in server directory) is the main configuration file.
- ▶ It configures
 - ▶ Project name
 - ▶ Admin password
 - ▶ Components
 - ▶ Annotator accounts

Components of WAT-SL

- ▶ Annotation components to describe jobs in the annotation interface
- ▶ 2 components of WAT-SL are
 - ▶ Text box (Used for comments)
 - ▶ Segment labeling
- ▶ Components can be configured project wide with `/project/<component-name>.conf`
- ▶ Components can be configured for each task that overrides settings in the project wide configuration file in `/tasks/<task-name>/<component-name>.conf`

Task 2 : Splitting speechtext to sentences

- Solution: Using `sent_tokenize` of NLTK to separate short text (speechtext) to list of sentences

In non-functional linguistics, a sentence is a textual unit consisting of one or more words that are grammatically linked. In functional linguistics, a sentence is a unit of written texts delimited by graphological features such as upper case letters and markers such as periods, question marks, and exclamation marks. This notion contrasts with a curve, which is delimited by phonologic features such as pitch and loudness and markers such as pauses.



[In non-functional linguistics, a sentence is a textual unit consisting of one or more words that are grammatically linked]

[In functional linguistics, a sentence is a unit of written texts delimited by graphological features such as upper case letters and markers such as periods, question marks, and exclamation mark]

[This notion contrasts with a curve, which is delimited by phonologic features such as pitch and loudness and markers such as pauses]

speechtext
<p>Mr. Speaker, on this last day of Women's History Month, I want to talk about a wonderful woman who inspired me a great deal.</p> <p>The theme of this year's Women's History Month was "Because of Her". Because of her, I fought my first political fight to prevent the closure of the Myrand ski hill, where I would go snowboarding. I was eight years old. Because of her, I fought my second and third political fights, with her in fact, against the forced municipal mergers. Because of her, I enjoyed Plage-Jacques-Cartier park during my entire childhood, and I still enjoy it today. Because of her, thousands of children back home in Sainte Foy have enjoyed affordable playgrounds. Many have also benefited from affordable housing because she thought it was better to invest in families than in bricks and mortar. She inspired me to get into politics. She knew how to navigate a man's world.</p> <p>I am talking about mayor Andrée P. Boucher. Unfortunately, she left us too soon. She was a generous, kind, and genuine woman to whom I owe a great debt, as do all citizens of Quebec City and Sainte Foy.</p> <p>Thank you, Mayor Boucher.</p>



sentences
mr. speaker, on this last day of women's history month, i want to talk about a wonderful woman who inspired me a great deal.
the theme of this year's women's history month was "because of her".
because of her, i fought my first political fight to prevent the closure of the myrand ski hill, where i would go snowboarding.
i was eight years old.
because of her, i fought my second and third political fights, with her in fact, against the forced municipal mergers.
because of her, i enjoyed plage-jacques-cartier park during my entire childhood, and i still enjoy it today.
because of her, thousands of children back home in sainte foy have enjoyed affordable playgrounds.
many have also benefited from affordable housing because she

- Version 1: Only keeping "maintopic", "speechtext" and adding ordering column

order	maintopic	sentences	#
1	statements by members	on this last day of women's history m...	
2	statements by members	the theme of this year's women's his...	
3	statements by members	because of her, i fought my first polit...	
4	statements by members	i was eight years old.	
5	statements by members	because of her, i fought my second ...	
6	statements by members	because of her, i fought my second ...	

Pros	Cons
Keep the size of .csv file as small as possible, only save the most necessary information	Have no information about which sentences come from which speechtext record

- Version 2: the same with Version 1, but adding "basepk" field to keep the information of speechtext that each sentences belong

	basepk	maintopic	sentences	#
10	4685947	statements by members	she knew how to navigate a man's world.	
11	4685947	statements by members	i am talking about mayor andrée p. bouc...	
12	4685947	statements by members	unfortunately, she left us too soon.	
13	4685947	statements by members	she was a generous, kind, and genuine ...	
14	4685947	statements by members	thank you, mayor boucher.	
15	4682434	statements by members	it is an honour to rise and mark world tea...	
16	4682434	statements by members	this day of recognition is an initiative put ...	
17	4682434	statements by members	this year's theme, "valuing teachers, impr...	
18	4682434	statements by members	that really is something we need to think ...	

Pros	Cons
Keep the size of .csv file small, keep the relationship between sentences and speechtext	no information about subtopic as well as speaker name

- Version 3: keep all field that we think it can be useful now and near future such as “basepk”, "speechdate", "main topic", "sub topic“, "speechtext“ because some field maybe will not be used any more in future such as "opid"

	basepk	speechdate	maintopic	subtopic	sentences	#
1	4685947	31/10/2016	statements by...	andrée p. bouc...	on this last day of...	
2	4685947	31/10/2016	statements by...	andrée p. bouc...	the theme of this ...	
3	4685947	31/10/2016	statements by...	andrée p. bouc...	because of her, i ...	
4	4685947	31/10/2016	statements by...	andrée p. bouc...	i was eight years ...	

Pros	Cons
only keep the necessary information	increase the size of .csv file

Statement by members	After spliting
Num of sentences	31559
Num of speechtext	4165
Average sentences each speechtext	7.58

Oral questions	After spliting
Num of sentences	19472
Num of speechtext	3660
Average sentences each speechtext	5.32

► Statement by members

	Origin	Version 1	Version 2	Version 3
Num of field	15	3	4	6
Memory (kb)	9121	4749	4996	5942
Proportion	100%	52,07%	54,77%	65,15%

► Oral questions

	Origin	Version 1	Version 2	Version 3
Num of field	15	3	4	6
Memory (kb)	5469	2536	2689	3173
Proportion	100%	46,37%	49,17%	58,02%

Approach 2 for the same task

- ▶ NLTK Punkt Sentence Tokenizer
- ▶ This tokenizer divides a text into a list of sentences, by using an unsupervised algorithm to build a model for abbreviation words, collocations, and words that start sentences.
- ▶ It must be trained on a large collection of plaintext in the target language before it can be used.
- ▶ The NLTK data package includes a pre-trained Punkt tokenizer for English.

Modification in the results of previous step

- ▶ Issue with quotes “This is an example. It explains how quotes should work.”
- ▶ NLTK Punkt Sentence Tokenizer output will be 2 sentences
 - ▶ “This is an example.
 - ▶ It explains how quotes should work.”
- ▶ Where as actual output the entire content within “ “ should be consider as one sentence
 - ▶ “This is an example. It explains how quotes should work.”

Results

- ▶ Punkt Sentence tokenizer is applied on controversial topics found in Oral questions and Statement by members.

	No of Speech	No of fragmented sentences
Oral Questions	3660	19685
Statement by members	4165	32106

Task 3 : Applying WAT-SL on Parliament data

- ▶ WAT-SL works with data in a txt file
- ▶ So fragmented speech from Oral questions and Statement

Controversial topics for Statement by members. (32106 segments remaining)

on this last day of Women's History Month, I want to talk about a wonderful woman who inspired me a great deal. ? The theme of this year's Women's History Month was ?Because of Her?. ? Because of her, I fought my first political fight to prevent the closure of the Myrand ski hill. ? Argumentative ?boarding. ? I was eight years old. ? Because of her, I fought my second and third political fights, with her in fact, against the forced municipalization of her, I enjoyed Plage-Jacques-Cartier park during my entire childhood, and I still enjoy it today. ? Non argumentative ? of her, I enjoyed Plage-Jacques-Cartier park during my entire childhood, and I still enjoy it today. ? Because of her, thousands of children back home in Sainte Foy have enjoyed affordable playgrounds. ? Many have also benefited from affordable housing because she thought it was better to invest in families than in bricks and mortar. ? She inspired me to get into politics. ? She knew how to navigate a man's world. ? I am talking about mayor Andrée P. Boucher. ? Unfortunately, she left us too soon. ? She was a generous, kind, and genuine woman to whom we owe a great debt, as do all citizens of Quebec City and Sainte Foy. ? Thank you, Mayor Boucher. ? it is an honour to rise and mark World Teachers' Day. ? This day of recognition is an initiative put forward by the United Nations Educational, Scientific and Cultural Organization, or UNESCO, and is held annually on October 5. ? The purpose of this day is appreciating, assessing, and improving the educators of the world. ? This year's theme, ?Valuing teachers, improving their status?, compels us to examine and resolve the problems that directly affect teachers, the people to whom we entrust our children's education. ? That really is something we need to think about because, all too often, the people who have dedicated themselves to this profession in Canada do not get the same level of respect as their counterparts elsewhere in the world. ? Thank goodness for French teachers. ? To the teachers in my community of London North

Annotation results

- Results are stored inside ‘Results’ directory for each completed task as a text file

```
segment-labeling.0 = anecdote
segment-labeling.1 = continued
segment-labeling.10 = assumption
segment-labeling.11 = nounit
segment-labeling.12 = continued
segment-labeling.13 = continued
segment-labeling.14 = assumption
segment-labeling.15 = nounit
segment-labeling.16 = continued
segment-labeling.17 = continued
segment-labeling.18 = anecdote
segment-labeling.19 = anecdote
segment-labeling.2 = continued
segment-labeling.20 = continued
segment-labeling.21 = anecdote
segment-labeling.22 = continued
segment-labeling.23 = continued
segment-labeling.24 = assumption
segment-labeling.25 = continued
segment-labeling.26 = continued
segment-labeling.27 = continued
segment-labeling.28 = continued
segment-labeling.29 = anecdote
segment-labeling.3 = continued
segment-labeling.30 = continued
segment-labeling.31 = continued
segment-labeling.32 = continued
segment-labeling.33 = anecdote
segment-labeling.34 = assumption
segment-labeling.35 = nounit
```