

How can we capture multiword expressions?

Seongmin Mun¹, Guillaume Desagulier², Anne Lacheret³ , Kyungwon Lee⁴

¹ Lifemedia Interdisciplinary Program, Ajou University, South Korea

^{1,3} UMR 7114 MoDyCo - CNRS, University Paris Nanterre, France

² UMR 7114 MoDyCo - University Paris 8, CNRS, University Nanterre

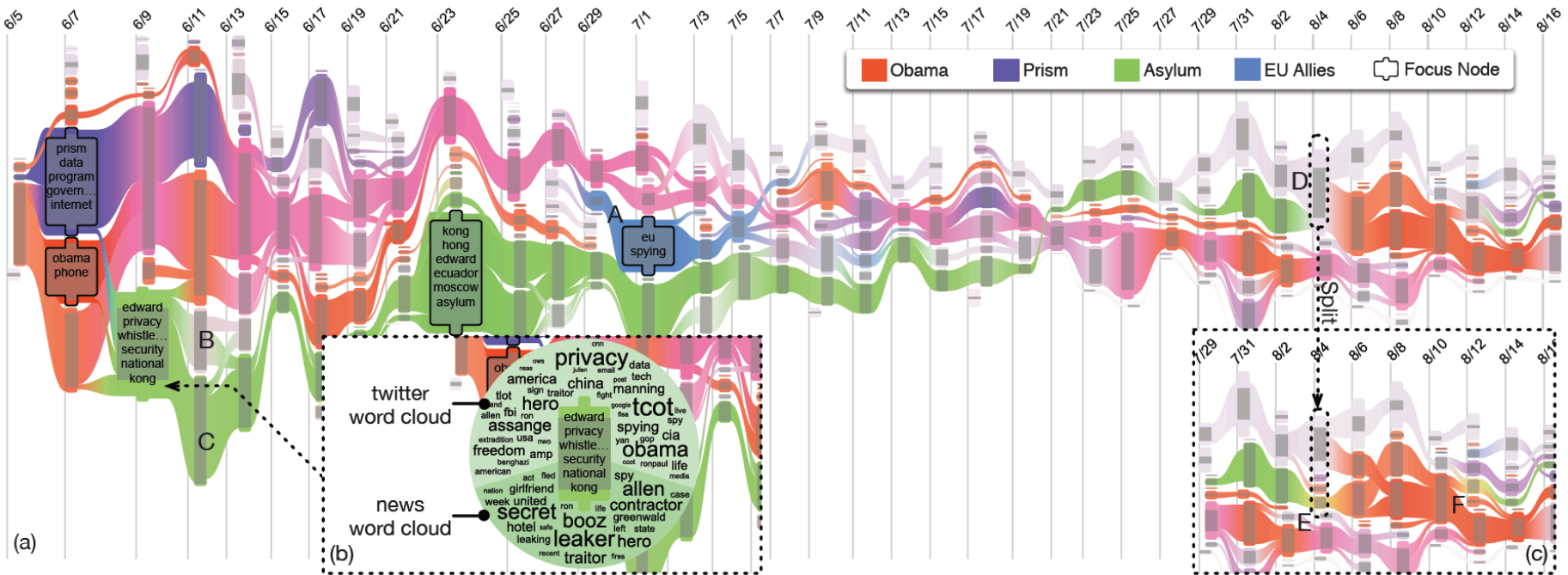
⁴ Department of Digital Media, Ajou University, South Korea

Introduction

Topics in a text corpus include features and information.

Analyzing these topics can improve a user's understanding of the corpus.

Previous studies



WEIWEI CUI SHIXIA LIU Z. W. H. W.: How hierarchical topics evolve in large text corpora. In *IEEE Transactions on Visualization and Computer Graphics* (2014), vol. 20, pp. 2281–2290.

Research background and purpose

Topics can be broadly divided into two categories.

Research background and purpose

“With profound gratitude and great humility, I accept your nomination for the presidency of the United States.”

Research background and purpose

“With profound *gratitude* and great humility, I accept your nomination for the presidency of the United States.”

Gratitude → meaning that can be expressed in one word

Research background and purpose

“With profound gratitude and great humility, I accept your nomination for the presidency of the *United States*.”

United States → meaning must be described using a combination of words.

Research background and purpose

How can we capture multiword expressions?

To this aim, we designed an algorithm.

Data processing



Data processing



Data processing

Raw corpus

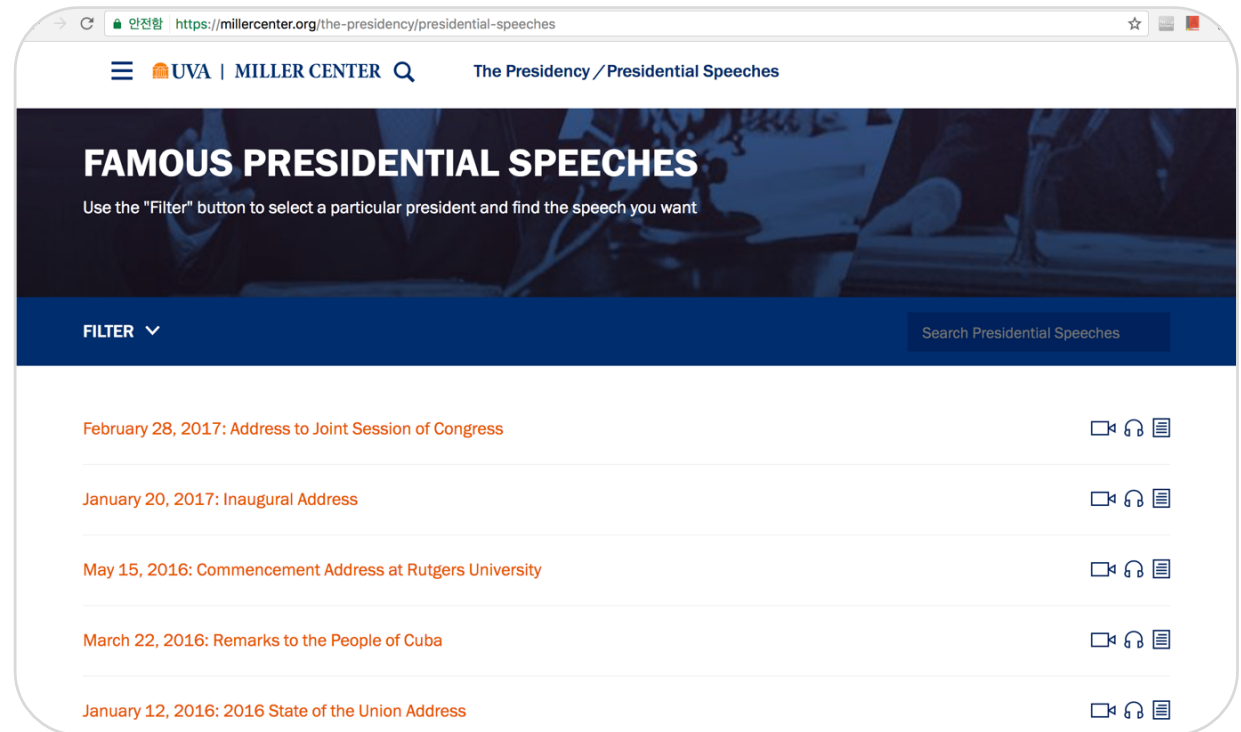
Processing

Topic
candidate

Topic
validation

Generate
topics

Raw corpus
(U.S. president speeches)



<https://millercenter.org/the-presidency/presidential-speeches>

Data processing

Raw corpus

Processing

Topic candidate

Topic validation

Generate topics

Raw corpus
(U.S. president speeches)

The screenshot displays the UVA Miller Center website at <https://millercenter.org/the-presidency/presidential-speeches>. The page features a header with the UVA Miller Center logo and a search bar. The main content area is titled "FAMOUS PRESIDENTIAL SPEECHES" and includes a subtext: "Use the 'Filter' button to select a particular president and find the speech you want". Below this, there is a "FILTER" dropdown menu and a "Search Presidential Speeches" button. A list of speeches is shown, with the first entry highlighted: "February 28, 2017: Address to Joint Session of Congress". Other visible entries include "January 20, 2017: Inaugural Address" and "May 15, 2016: Commencement Address at Rutgers University".

The browser's developer tool is open on the right side, showing the "Elements" panel. The selected element is a link with the href `/the-presidency/presidential-speeches/february-28-2017-address-joint-session-congress`. The "Styles" panel shows the default styling for the link, including `text-decoration: none;`. The "Console" panel is also visible at the bottom of the developer tool.

Data processing

Raw corpus

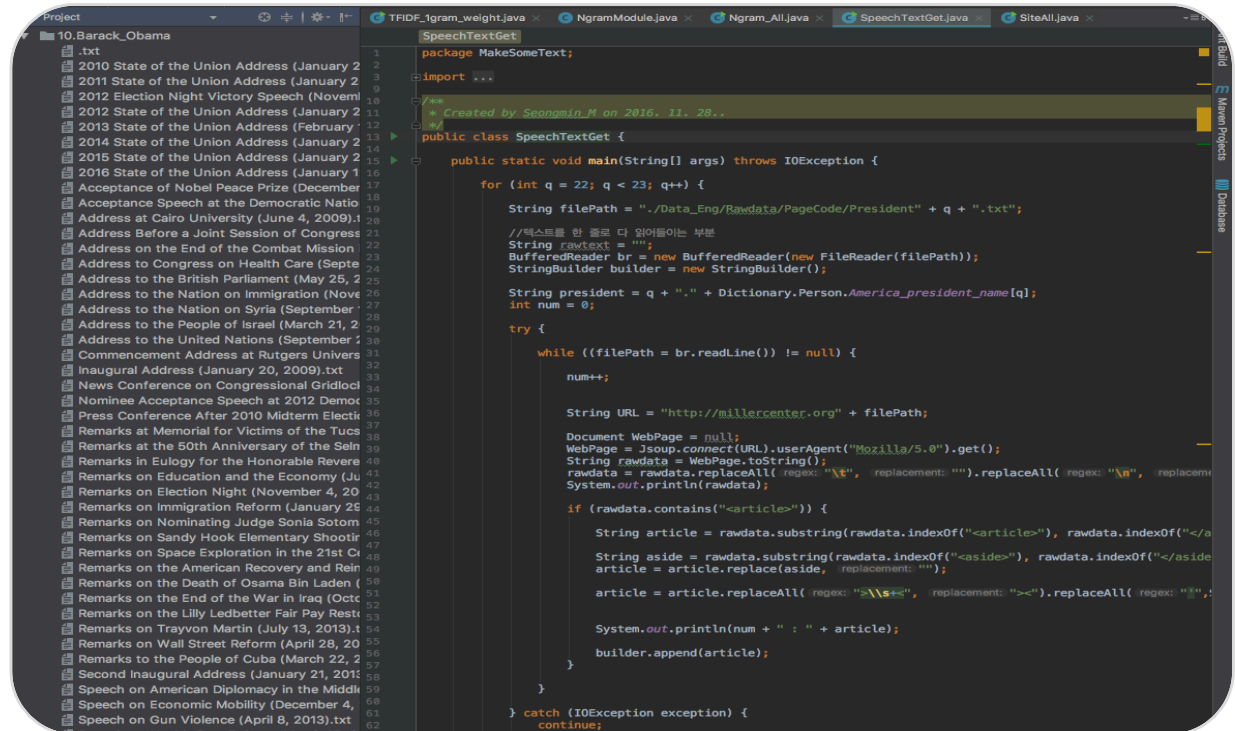
Processing

Topic candidate

Topic validation

Generate topics

Raw corpus
(U.S. president speeches)



```
Project
├── 10.Barack_Obama
│   ├── .txt
│   ├── 2010 State of the Union Address (January 20, 2010).txt
│   ├── 2011 State of the Union Address (January 20, 2011).txt
│   ├── 2012 Election Night Victory Speech (November 6, 2012).txt
│   ├── 2012 State of the Union Address (January 20, 2012).txt
│   ├── 2013 State of the Union Address (February 12, 2013).txt
│   ├── 2014 State of the Union Address (January 20, 2014).txt
│   ├── 2015 State of the Union Address (January 20, 2015).txt
│   ├── 2016 State of the Union Address (January 20, 2016).txt
│   ├── Acceptance of Nobel Peace Prize (December 10, 2009).txt
│   ├── Acceptance Speech at the Democratic National Convention (September 12, 2008).txt
│   ├── Address at Cairo University (June 4, 2009).txt
│   ├── Address Before a Joint Session of Congress (January 20, 2009).txt
│   ├── Address on the End of the Combat Mission in Iraq (May 2, 2003).txt
│   ├── Address to Congress on Health Care (September 9, 2009).txt
│   ├── Address to the British Parliament (May 25, 2009).txt
│   ├── Address to the Nation on Immigration (November 20, 2009).txt
│   ├── Address to the Nation on Syria (September 27, 2001).txt
│   ├── Address to the People of Israel (March 21, 2009).txt
│   ├── Address to the United Nations (September 12, 2009).txt
│   ├── Commencement Address at Rutgers University (May 12, 2009).txt
│   ├── Inaugural Address (January 20, 2009).txt
│   ├── News Conference on Congressional Gridlock (January 22, 2009).txt
│   ├── Nominee Acceptance Speech at 2012 Democratic National Convention (September 5, 2012).txt
│   ├── Press Conference After 2010 Midterm Elections (November 4, 2010).txt
│   ├── Remarks at Memorial for Victims of the Tuskegee Army (September 11, 2009).txt
│   ├── Remarks at the 50th Anniversary of the Selma to Montgomery March (March 3, 2009).txt
│   ├── Remarks on Education and the Economy (July 13, 2013).txt
│   ├── Remarks on Election Night (November 4, 2008).txt
│   ├── Remarks on Immigration Reform (January 26, 2009).txt
│   ├── Remarks on Nominating Judge Sonia Sotomayor (September 8, 2009).txt
│   ├── Remarks on Sandy Hook Elementary Shooting (December 20, 2012).txt
│   ├── Remarks on Space Exploration in the 21st Century (September 12, 2008).txt
│   ├── Remarks on the American Recovery and Reinvestment Act (February 18, 2009).txt
│   ├── Remarks on the Death of Osama Bin Laden (May 1, 2011).txt
│   ├── Remarks on the End of the War in Iraq (October 2, 2003).txt
│   ├── Remarks on the Lilly Ledbetter Fair Pay Restatement (July 27, 2009).txt
│   ├── Remarks on Trayvon Martin (July 13, 2013).txt
│   ├── Remarks on Wall Street Reform (April 28, 2009).txt
│   ├── Remarks to the People of Cuba (March 22, 2009).txt
│   ├── Second Inaugural Address (January 21, 2013).txt
│   ├── Speech on American Diplomacy in the Middle East (September 12, 2009).txt
│   ├── Speech on Economic Mobility (December 4, 2009).txt
│   ├── Speech on Gun Violence (April 8, 2013).txt
│   └── ...
└── ...

TFIDF_1gram_weight.java
NgramModule.java
Ngram_All.java
SpeechTextGet.java
SiteAll.java

SpeechTextGet.java
package MakeSomeText;
import java.io.*;
import java.net.*;
import java.util.*;
import java.util.regex.*;

// Created by Seonmin.M on 2016. 11. 28..
public class SpeechTextGet {
    public static void main(String[] args) throws IOException {
        for (int q = 22; q < 23; q++) {
            String filePath = "../Data_Eng/Rawdata/PageCode/President" + q + ".txt";
            // 텍스트를 한 줄로 다 읽어들이는 부분
            String rawText = "";
            BufferedReader br = new BufferedReader(new FileReader(filePath));
            StringBuilder builder = new StringBuilder();

            String president = q + "." + Dictionary.Person.America_president_name[q];
            int num = 0;

            try {
                while ((filePath = br.readLine()) != null) {
                    num++;

                    String URL = "http://millercenter.org" + filePath;
                    Document WebPage = null;
                    WebPage = Jsoup.connect(URL).userAgent("Mozilla/5.0").get();
                    String rawdata = WebPage.toString();
                    rawdata = rawdata.replaceAll(regex: "\\t", replacement: "").replaceAll(regex: "\\n", replacement: "");
                    System.out.println(rawdata);

                    if (rawdata.contains("<article>")) {
                        String article = rawdata.substring(rawdata.indexOf("<article>"), rawdata.indexOf("</article>"));
                        String aside = rawdata.substring(rawdata.indexOf("<aside>"), rawdata.indexOf("</aside>"));
                        article = article.replace(aside, "");
                        article = article.replaceAll(regex: "<\\s*>", replacement: ">").replaceAll(regex: "<\\s*>", replacement: ">");

                        System.out.println(num + " : " + article);
                        builder.append(article);
                    }
                }
            } catch (IOException exception) {
                continue;
            }
        }
    }
}
```

Data processing



Data processing



Processing

- N-grams
- POS tagging

Pre-processing

- Cleaning with RegExp
- Lemmatization
- Tokenization
- Lowercasing

N-gram method is a contiguous sequence of ***N*** items from a given sequence of text.

Data processing



Processing

- N-grams
- POS tagging

Pre-processing

- Cleaning with RegExp
- Lemmatization
- Tokenization
- Lowercasing

“Time flies like an arrow.”

Data processing



Processing

- N-grams
- POS tagging

Pre-processing

- Cleaning with RegExp
- Lemmatization
- Tokenization
- Lowercasing

“Time flies like an arrow.”

Unigram : Time, flies, like, an, arrow.

Bigram : Time flies, flies like, like an, an arrow.

Trigram : Time flies like, flies like an, like an arrow.

Data processing



Processing

- N-grams
- POS tagging

Pre-processing

- Cleaning with RegExp
- Lemmatization
- Tokenization
- Lowercasing

```
2.George_W._Bush_3gram_result.txt x
1 number,text,count
2 1,the united states,128
3 2,men and women,73
4 3,the middle east,73
5 4,the american people,70
6 5,and we will,65
7 6,of the world,43
8 7,in the middle,41
9 8,one of the,39
10 9,in the world,39
11 10,weapons of mass,38
12 11,members of congress,38
13 12,and that is,38
14 13,i want to,38
15 14,it is the,36
16 15,the united nations,36
17 16,of our country,34
18 17,of the united,34
19 18,a lot of,34
20 19,thank you for,33
21 20,ask you to,33
22 21,is going to,32
23 22,of mass destruction,32
24 23,i ask you,32
25 24,want to thank,31
26 25,around the world,30
```

```
1.Barack_Obama_2gram_result.txt x
1 number,text,count
2 1,of the,802
3 2,in the,787
4 3,to the,429
5 4,that is,421
6 5,of our,402
7 6,it is,379
8 7,and the,378
9 8,we have,375
10 9,for the,336
11 10,we can,319
12 11,that we,316
13 12,we will,305
14 13,to be,299
15 14,on the,289
16 15,the world,280
17 16,going to,256
18 17,we are,251
19 18,and i,240
20 19,is not,229
21 20,that the,221
22 21,want to,213
23 22,will be,212
24 23,the united,212
25 24,and we,212
26 25,is the,207
```

Data processing



Data processing



Topic candidate extraction & filtering

- Frequency counting
- Filters :
 - ✓ Stopwords
 - ✓ Thresholds

```
public class Remove {  
    public static String[] stopwords_three = {  
        "it",  
        "is",  
        "we",  
        "are",  
        "this",  
        "be",  
        "may",  
        "would",  
        "am",  
        "more",  
        "than",  
        "do",  
        "that",  
        "can",  
        "not",  
        "could",  
        "sould",  
        "shall",  
        "will",  
        "were",  
        "was",  
        "might",  
        "all",  
        "so",  
        "you",  
        "he",  
        "him",  
        "his",  
        "she",  
        "her",  
        "your",  
        "me",  
    }  
}
```

```
public static String[] stopwords = {  
    "t",  
    "st",  
    "de",  
    "ss",  
    "el",  
    "ho",  
    "em",  
    "men",  
    "ere",  
    "ad",  
    "la",  
    "pro",  
    "fe",  
    "wit",  
    "vi",  
    "ted",  
    "eve",  
    "iv",  
    "era",  
    "ear",  
    "va",  
    "ive",  
    "led",  
    "owe",  
    "tho",  
    "gi",  
    "a",  
    "will",  
    "able",  
    "about",  
    "above",  
    "abst",  
}
```

Data processing



Data processing



Topic validation

- Human annotation
- Matching with **Dictionaries**

English dictionaries

- THE DEVIL'S DICTIONARY ((C)1911 Released April 15 1993)
- Easton's 1897 Bible Dictionary
- Elements database 20001107
- The Free On-line Dictionary of Computing (27 SEP 03)
- U.S. Gazetteer (1990)
- The Collaborative International Dictionary of English v.0.44
- Hitchcock's Bible Names Dictionary (late 1800's)
- Jargon File (4.3.1, 29 June 2001)
- Virtual Entity of Relevant Acronyms (Version 1.9, June 2002)
- WordNet (r) 2.0
- CIA World Factbook 2002
- User Dictionary


Data processing



Visual system

Source : President's speeches in U.S.

George_Washington



Term

-1st President of the United States (1789 - 1797)

Born

-February 22, 1732, Pope's Creek, Virginia

Political Party

-Federalist

Died

-December 14, 1799

Speech transcript (Number of texts : 21)

First Inaugural Address (April 30, 1789)

George Washington

Transcript

Fellow Citizens of the Senate and the House of Representatives:
Among the vicissitudes incident to life, no event could have filled me with greater anxieties than that of which the notification was transmitted by your order, and received on the fourteenth day of the present month. On the one hand, I was summoned by my Country, whose voice I can never hear but with veneration and love, from a retreat which I had chosen with the fondest predilection, and, in my flattering hopes, with an immutable decision, as the asylum of my declining years: a retreat which was rendered every day more necessary as well as more dear to me, by the addition of habit to inclination, and of frequent interruptions in my health to the gradual waste committed on it by time. On the other hand, the magnitude and difficulty of the trust to which the voice of my Country called me, being sufficient to awaken in the wisest and most experienced of her citizens, a distrustful scrutiny into his qualification, could not but overwhelm with dispondence, one, who, inheriting inferior endowments from nature and unpractised in the duties of civil administration, ought to be peculiarly conscious of his own deficiencies. In this conflict of emotions, all I dare aver, is, that it has been my faithful study to collect my duty from a just appreciation of every circumstance, by which it might be affected. All I dare hope, is, that, if in executing this task I have been too much swayed by a grateful remembrance of

1.Select N-gram

Uni-gram

Bi-gram

Tri-gram

2.Select issue

All_speech

Economic_Growth

Social_Welfare

US_Policies_War

Health_Care

Immigration

Humanitarian_Aid

Protecting_US_Terror

Establishing_Democracy

Promoting_US_Strength

3.Select president

George_Washington

John_Adams

Thomas_Jefferson

James_Madison

James_Monroe

John_Quincy_Adams

Andrew_Jackson

Martin_Van_Buren

William_Harrison

John_Tyler

James_Polk

Zachary_Taylor

Topic information

Please select topic in the Word Cloud...

Color of topics

Frequency(0-10)

Frequency(11-20)

Frequency(21-30)

Frequency(31-40)

Frequency(41-50)

Frequency(51-60)

Word Cloud with multiword

congress

policy

course

military

interest

national

militia

constitution

great

parts

regard

treaty

peace

object

full

state

hope

desire

good

nation

duty

force

future

laws

time

law

general

public

acts

best

government

trust

laid

country

union

day

place

people

war

proper

duties

senate

foreign

powers

treaties

liberty

happiness

spirit

power

reason

Sorted result

1 government(103)

2 public(65)

3 country(64)

4 nation(55)

5 laws(54)

6 time(44)

7 peace(43)

8 people(42)

9 general(40)

10 constitution(39)

11 good(34)

12 union(33)

13 state(33)

14 war(33)

15 national(32)

16 congress(32)

17 foreign(31)

18 proper(31)

19 law(30)

20 powers(30)

21 great(30)

22 treaty(29)

23 duty(28)

24 justice(27)

© 2017, All rights reserved. (latest corpus uploading : 07/01/2017)

<http://ressources.modyco.fr/sm/MultiwordVis/>

Ambiguous sentence

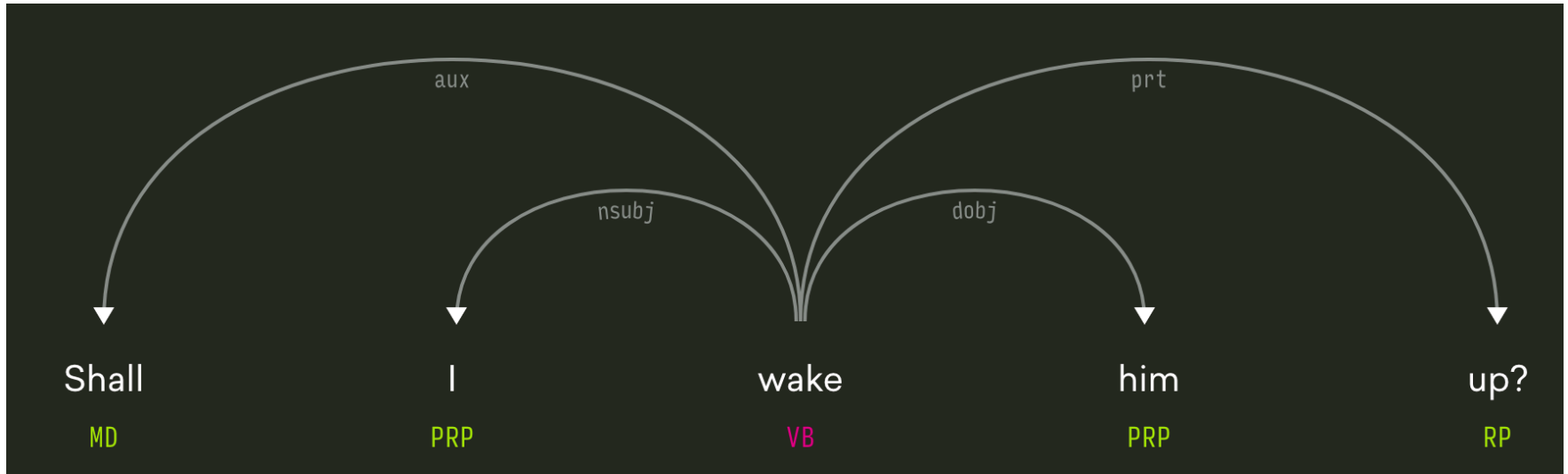
“Shall I wake him up?”

Ambiguous sentence

“Shall I **wake** him **up**?”

We can't extract wake up if we only use N-gram algorithm.

Dependency tag



Dependency tag can provide a simple description of the grammatical relationships in a sentence.

Improving algorithm

Result of dependency graph below

dependency graph:

- > wake/VBP (root)
 - > Shall/NNP (nsubj)
 - > I/PRP (dep)
 - > him/PRP (dobj)
 - > up/RP (compound:prt)
 - > ?/. (punct)

Result of multiword candidates

wake Shall
Shall I
wake Shall I
wake him
wake up
wake ?

Improving algorithm

Final result below

```
0. wake is meaningful : wake
1. shall is meaningful : shall
2. i is meaningful : i
3. up is meaningful : up
4. shall i is meaningful : shall i
5. him is meaningful : him
```

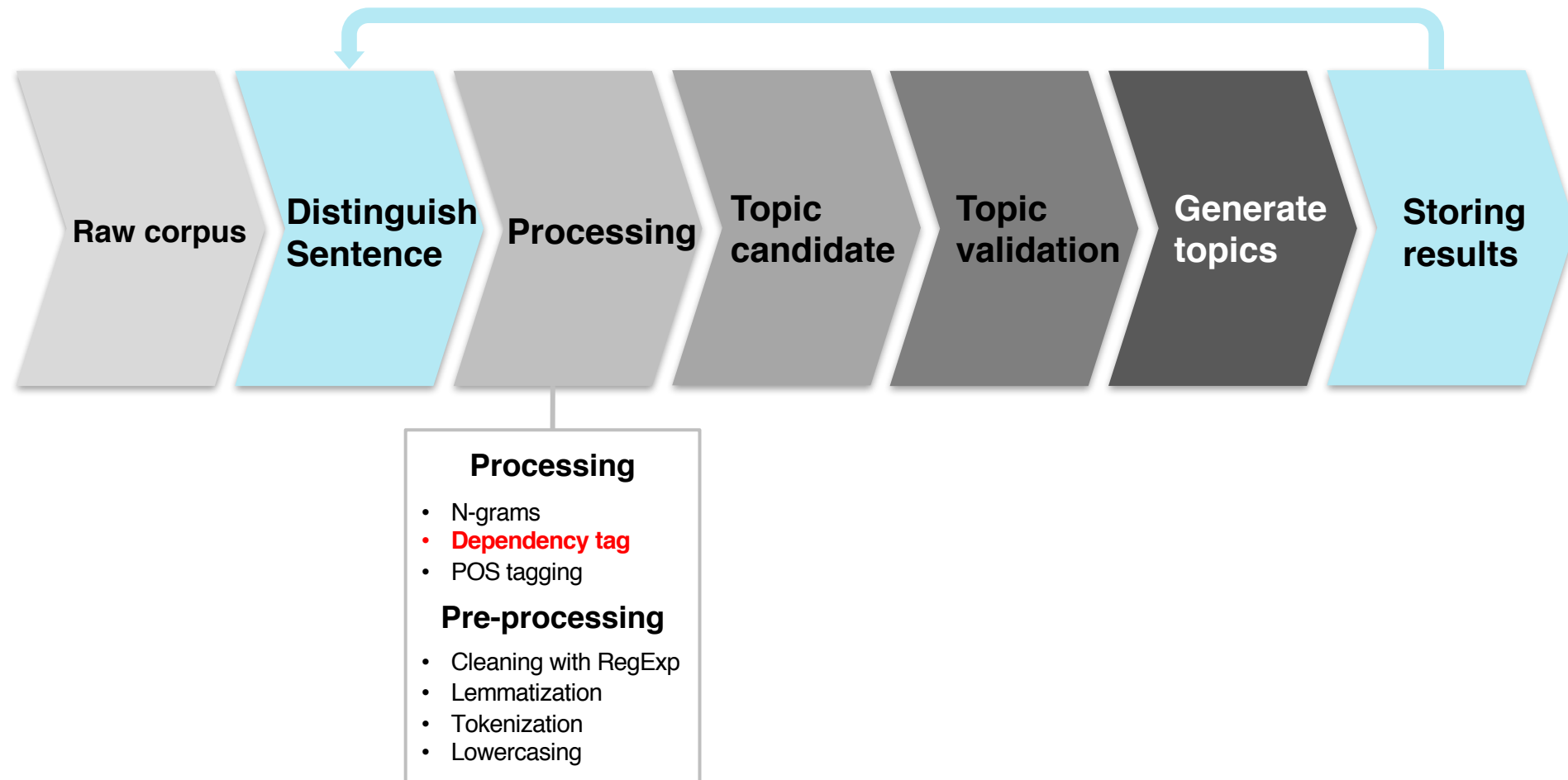
N-gram

Final result below

```
0. wake is meaningful : wake
1. shall i is meaningful : shall i
2. i is meaningful : i
3. wake up is meaningful : wake up
4. up is meaningful : up
5. him is meaningful : him
6. shall is meaningful : shall
```

Dependency tag

Data processing



Q&A

Thank you for listening.
stat34@ajou.ac.kr