How can we capture multiword expressions?

**Seongmin Mun**

1. Lifemedia Interdisciplinary Program, Ajou University, South Korea
2. UMR 7114 MoDyCo - CNRS, University Paris Nanterre, France
3. 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

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

- Raw corpus
- Processing
- Topic candidate
- Topic validation
- Generate topics
Data processing

- Raw corpus
- Processing
- Topic candidate
- Topic validation
- Generate topics
Data processing

Raw corpus
(U.S. president speeches)

https://millercenter.org/the-presidency/presidential-speeches
Data processing

Raw corpus
(U.S. president speeches)
Data processing

Raw corpus
(U.S. president speeches)
Data processing

- Raw corpus
- Processing
- Topic candidate
- Topic validation
- Generate topics
Data processing

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

Processing
- N-grams
- POS tagging

Pre-processing
- Cleaning with RegExp
- Lemmatization
- Tokenization
- Lowercasing
Data processing

“Time flies like an arrow.”

Processing
- N-grams
- POS tagging

Pre-processing
- Cleaning with RegExp
- Lemmatization
- Tokenization
- Lowercasing
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
Data processing

Raw corpus → Processing → Topic candidate → Topic validation → Generate topics
Data processing

Topic candidate extraction & filtering
- Frequency counting
- Filters:
  - Stopwords
  - Thresholds
Data processing

- Raw corpus
- Processing
- Topic candidate
- Topic validation
- Generate topics
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

- Raw corpus
- Processing
- Topic candidate
- Topic validation
- Generate topics
Visual system

Source: President's speeches in U.S.

George Washington

First Inaugural Address (April 30, 1789)

George Washington

Transcript

Fellow Citizens of the Senate and the House of Representatives:
Among the vexations 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 this 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 flitting hopes, with an immutable decision, as the eum to my declining years: a retreat which was waded 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 widest and most experienced of her citizens, a distrustful scrutiny into his qualification, could not but overwhelm with dissonance, one, whinching inferior endorsements 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, in executing this task I have been too much swayed by a grateful remembrance of

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 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

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

N-gram

Dependency tag
Data processing

- Raw corpus
- Distinguish Sentence
- Processing
- Topic candidate
- Topic validation
- Generate topics
- Storing results

Processing
- N-grams
- Dependency tag
- POS tagging

Pre-processing
- Cleaning with RegExp
- Lemmatization
- Tokenization
- Lowercasing
Thank you for listening.
stat34@ajou.ac.kr