Frequency, Collocation, and Statistical Modeling of Lexical Items: A Case Study of Temporal Expressions in an Elderly Speaker Corpus

Sheng-Fu Wang¹, Jing-Chen Yang¹, Yu-Yun Chang¹, Yu-Wen Liu², Shu-Kai Hsieh¹

¹National Taiwan University
²National Taiwan Normal University

@ROCLING XXIII, 2011-SEP-09
Outline

- Introduction
- Methodology
- Statistical modeling
- Discussion
- Conclusion
• Aging caused not only physiological changes but also effects on language production for elderly people.
• The phenomenon of gerontology could be studied through discourse analysis. (Green, 1993)

• The social roles of elderly people may be embedded in the conversation when they share personal experience or judgment of the past and the present. (Kuo, 2008)
  ○ Temporal expressions might pervade as the anchoring points in the conversation and reveal the speech behavior of elderly people.
Introduction

- Statistical modeling can serve to describe a given set of data, and may help us develop NLP-oriented lexicographic modules.

- Two kinds of data input in modeling
  - Frequency distribution pattern:
    - e.g. word frequency, variability of frequency, n-gram distribution
  - Collocational distribution pattern
    - e.g. dependency with syntactic patterns, co-occurrence with particular word type
• **Purpose of the study:**

  - A preliminary analysis of the use of temporal expressions in the elderly speakers’ speech.
  - Methodological comparison and contrast of different input data types and their results.
  - To provide a tool for researchers to make a quick preview of their data.
Methodology: The corpus

- The elderly Mandarin speaker corpus
  - Four recordings from four pairs of elderly people
  - Each pair consists of one male and one female.
  - Native speakers of Mandarin and Taiwanese Southern Min
  - Mean age: 65.75 years old (SD = 6.16)
  - Face-to-face conversation about life in the past and the present
  - Total length: 2 hr 25 min
Methodology: Corpus information

- Segmentation and POS tagging are based on the standards of Sinica corpus.
- Code-switching occurred in the conversation, but is not included in the following analyses.

- Elderly people’s production in Mandarin
  - 3,739 intonation units
  - 18,076 word tokens
Methodology: tool - R

- Use R to do
  - Preprocessing
  - Word frequency calculation
  - KWIC (KeyWord In Context) extraction
  - Statistical modeling

```r
> test <- kwic("現在\(\text{Nd}\)\"", 3, 3, cleanfile13)
> test[[27]]
[1] "381\text{tSF}:\text{t以後(Nd) 現在(Nd) 坐(VA) 飛機(Na) 你(Nh) 不用(D) 怕(VK)""
[2] "382\text{tSF}:\text{t我(Nh)}"
[3] "383\text{tSF}:\text{t以前(Ng) 也(D) 不(D) 敢(VL) 坐(VA)"
[4] "384\text{tSF}:\text{t我(Nh) 現在(Nd) 他(Nh(T)"
[5] "385\text{tSF}:\text{t這(Nep) 個(Nf) 坐(VA) 飛機(Na) 也(D) 很(Dfa) 厲害(VH)"
[6] "386\text{tSF}:\text{t捷運(Na) 真的(D) 很(Dfa) 方便(VH)"
[7] "387\text{tSF}:\text{t捷運(Na) 也(D) 很(Dfa) 厲害(VH)"
```
Temporal expressions

- **Present:** 現在(now), 最近(recently)
- **Past:** 以前(before), 小時候(in one’s childhood), 民國(R.O.C. year), 當初(back then)

| Term                  | Freq | Rank |
|-----------------------|------|------|
| 現在(now)             | 169  | 1    |
| 以前(before)          | 169  | 1    |
| 小時候(in one’s childhood) | 12  | 2    |
| 民國(R.O.C. year)     | 11   | 3    |
| 當初(back then)       | 9    | 4    |
| 最近(recently)        | 6    | 5    |
Divisive hierarchical clustering
- A group of entities is first divided into large groups and then smaller groups are classified.
- Better at classifying entities into large clusters (Baayen, 2008)
- Nice visualization

To examine whether the terms for “the present” and “the past” can really be grouped into clusters different in temporality.
Input types: Different dimensions of data on temporal terms

- **Frequency distribution pattern**
  - Frequencies in texts as variables
  - Frequencies in subsets of different sizes as variables
    - Subset size: 10, 50, 200, 500 words

- **Collocational distribution pattern**
  - Collocational patterns with
    - All words (span = 3)
    - Particles: lah, hoNh, ah, oh, le, haNh, hioh, mah
  - Coll.strength (Gries, Hampe, & Schönefeld, 2005; Gries, 2007)
With frequencies in texts

word.dist[1]
diana (*, "NA")
With Subsets of small sizes

with frequencies in subsets of a size of 10

with frequencies in subsets of a size of 50

word.dist[[2]]
diana(*, "NA")

word.dist[[3]]
diana(*, "NA")
With Subsets of larger sizes

with frequencies in subsets of a size of 200

with frequencies in subsets of a size of 500

word.dist[4]
diana (*, "NA")

word.dist[5]
diana (*, "NA")
With Raw association/collection patterns
with Coll.strength
Evaluating different methods

- Evaluation with hierarchical clustering!
  - Input: the clustering table produced by R’s “cutree” function (in the “cluster” package)
Comparing clustering methods

Frequencies across parts of corpus

Collocation and association analysis

By text analysis is differentiated first

The statistically-based are differentiated first
Core expressions of the present and the past have very similar distributional patterns.

- Elderly speakers tend to compare the present with the past in the same textual domains.
Discussion: What do different input data tell us?

- “Input type” as a factor behind different results:
  - Frequencies in subsets produce similar clustering results
    - Revealing how these terms co-occur
  - By-text analysis is distinguished in the frequency cluster
    - Preferred usage of temporal expressions for different speakers or conversation events
Discussion: What do different input data tell us? (cont.)

- Collocation/association analyses constitute another major cluster
  - Reflecting what kinds of words these terms attract

- Statistical analyses on collocates stand out in the collocation/association cluster
  - Genuine association/collection patterns

- The prediction accuracy should be:
  - coll.strength > collocation pattern > frequency in subset > frequency in text
Conclusion

- This study demonstrates how the models make a quick preview of the data pattern, based on which researchers can move on to do a detailed analysis such as qualitative analysis.

- We suggest using coll.strength and collocation pattern as pre-analyzing tool to preview the data.
Thank you!