Personal Experience of Data Sharing in Text Analysis

Chris Khoo
Wee Kim Wee School of Communication & Information
Nanyang Technological University, Singapore
Why share data?

1. Retirement is looming
   - What impact have I had on my field, except for a few papers?

2. NTU is #1 in Asia
   - The best universities share resources with the research community
   - A way of providing leadership and supporting progress in the research community

3. Share resources with public to support educational use (and recreational and industry use)

4. Support intelligent Web
   - Semantic web and linked data applications
Resources in text mining and natural language processing

- **Lexical resources:** lists of words/phrases with semantic categorization
  - Sentiment lexicon
  - List of causal verbs

- **Lists of word patterns (regular expression patterns) for sentence categorization or extracting pieces of information**
  - Word patterns for extracting cause-effect information in text
  - Discourse markers for identifying rhetorical & metadiscourse functions (in genre analysis)
Resources in text mining and natural language processing (cont.)

- **Annotated corpora**
  - Social media data (product and drug reviews) annotated with sentiment categories
  - Journal articles annotated with rhetorical and metadiscourse functions, as well as information types, citation types and argument types

- **Text processing software**
  - Text segmentation (to analyse structure of documents and sentences)
  - Syntactic/semantic parsing (to identify relations between words or concepts)
WKWSCI Sentiment Lexicon

- Manually coded by WKWSCI undergrad students over a period of 4 years
  
  Total # words: ~ 30,000 words
  # positive words: ~ 3,000
  # negative words: ~ 7,000
  # neutral words: ~ 20,000

  *Includes adjectives, adverbs, verbs, nouns*

- Coded on a 7-point sentiment strength scale
  - Very positive, Positive, Slightly Positive
  - Neutral
  - Very negative, Negative, Slightly negative
Evidence Based Academic Writing Assistant

Title

Lexicon-Based Sentiment Analysis: Comparative Evaluation of Six Sentiment Lexicons

Introduction:
The core of the Introduction section is the research objective(s) or research question(s) statement. The text before the research objective introduces the broader research area, tradition, theory or context in which the study is situated. The research area is gradually narrowed down to the research problem/issue that the study addresses, including the specific research gap that the research objective addresses. The text will also indicate why the research issue is central, important or worth studying.

Introduction - Typical Structure

Move 1: Introduce the field
- Step 1: Introduce the research area/topic
- Step 2: Make a general statement about the area/topic (optional)
- Step 3: Outline the historical development of the area/topic (optional)
- Step 4: Elaborate on the research area/topic
- Step 5: Narrow down the topic (optional)

Move: Introduce the field
Step: Introduce the research area/topic

Description: Stating the topic of the current study.

Text Pattern(s):
1. researchers: research ... *investigate
2. extent: existing *literature ... *investigate
3. review: survey ... *investigate
4. body of literature
5. body of *study
6. emergence of
Example Sentences - Google Chrome

Introduction:
The core of the Introduction section is the research objective(s) or research question(s) statement. The text before the research objective introduces the broader research area, tradition, theory or context in which the study is situated. The research area is gradually narrowed down to the research problem/issue that the study addresses, including the specific research gap that the research objective addresses. The text will also indicate why the research issue is central, important or worth studying.

Introduction - Typical Structure

Move 1: Introduce the field

- Step 1: Introduce the research area/topic
- Step 2: Make a general statement about the area/topic (optional)
- Step 3: Outline the historical development of the area/topic (optional)
- Step 4: Elaborate on the research area/topic
- Step 5: Narrow down the topic (optional)

Move: Introduce the field

Step: Introduce the research area/topic

Description: Stating the topic of the current study.

Text Pattern(s):
1. researchers/research ... *investigate
2. extent/existing *literature ... *investigate
3. review/survey ... *investigate
4. body of literature
5. body of *study
6. emergence of
Questionnaire survey data

- Survey of the types of information shared on social networking sites by university students

Across 7 universities
Nanyang Technological University, Singapore
University of Malaya, Malaysia
University of Pondicherry, India
National Taiwan Normal University, Taiwan
Kyushu University, Japan
Mahasarakham University, Thailand
Keimyung University, South Korea
IRB application for questionnaire survey data

What will happen to the data after research completion?

- The data will be kept permanently in a secure room by the investigators for future research use.
- In addition, aggregated data (with no subject identification information) will be uploaded to NTU digital repository for public access.
Challenges

- Funding for resource preparation
  - Development, cleaning, formatting, documentation

- IP and copyright
  - Text analysis resources are usually built on more primitive resources which may be copyrighted or have IP issues

- IRB application
  - Is painful
  - What is acceptable to the review committee?

- Resources to support reuse of the data
  - Simple charting and analysis tools for survey data
  - Software for using the resource
Thank you very much!