Construction of an Efficient Hierarchical Semantic Frame Network using a Thesaurus

4th International Workshop on Generative Approaches to the Lexicon (GL2007)
Paris, 11th May, 2007

Toshiyuki Kanamaru†, Masaki Murata†,
and Hitoshi Isahara†

†National Institute of Information and Communication Technology (NICT)
Contents

- Motivation
  - FOCAL: theoretical framework for semantic analysis and annotation.
  - Method for efficiently discovering the semantic frame that is target of FOCAL analysis.

- Method
  - Use hierarchical structure of thesaurus

- Result and Discussion
  - Discover new semantic frames and efficiently analyze them

- Summary
Motivation

■ Background
  ■ FOCAL: The Frame-oriented Concept Analysis of Language proposed by Kuroda and Isahara.

■ Current procedure
  ■ Manual analysis of data from early stages necessary

■ What’s new in our field?
  ■ Efficient method for discovering semantic frame that is target of FOCAL analysis
What is FOCAL?

- **FOCAL (Frame-Oriented Concept Analysis):** theoretical framework for semantic analysis and annotation (Kuroda et al. 2004)
  - Frames characterized as “idealized model of situations”

- **FOCAL identified senses of Japanese verb “osou”** (attack, hit, strike, etc…) as related to 15 different situations/frames.
  - Claim was confirmed through psychological experiments (Nakamoto et al. 2005).
15 FOCAL frames with groups

- **G1**
  - F01 harm to Y caused by conflict between groups X and Y
  - F02 harm to Y caused by X’s invasion
  - F03 harm to Y caused by X’s robbery
  - F04 harm to Y caused by X’s violence
  - F05 harm to Y caused by X’s raping

- **G2**
  - F06 harm to Y caused by X’s preying attack
  - F07 harm to Y caused by X’s non-preying attack (e.g., X’s defense)

- **G3**
  - F08 harm to Y due to an unexpected accident X
  - F09 harm to Y caused by a natural phenomenon X (on a smaller scale, e.g., gust)

- **G4**
  - F10 harm to Y caused by a natural phenomenon X (on a larger scale, e.g., earthquake, flood)
  - F11 harm to Y caused by a natural phenomenon X (on a larger scale, e.g., spread of an epidemic)

- **G5**
  - F12 harm to Y caused by a social phenomenon X

- **F13** harm to Y caused by a disease X (non-temporary, e.g., cancer)
- **F14** harm to Y caused by a disease symptom X (temporary, e.g., heart attack)
- **F15** harm to Y caused by a bad feeling X (temporary, e.g., drowsiness)

* The red frames are discovered in newly version of analysis.
Current procedure

1. All usage of verb “osou” extracted from bilingual corpus using KWIC tools.
2. Following information was annotated manually to specify semantic frames of usages.
   - name of frame expressed in sentence S
   - subject phrase s of sentence S, and frame element of s
   - object phrase o of sentence S, semantic type of o, and frame element of o
   - semantic frame of sentence S
Problems with procedure

- All extracted usages are analyzed.
  - Many verbs have thousands of usages. Procedure that extracts all usages for analysis has limitations.

- Not oriented to getting semantic frame in proper perspective.
  - Specifying semantic frame for extracted usages is difficult until more analysis has been done. Now clear that semantic frames cannot be specified until the data has been partially analyzed.
Proposed method

1. Text data collected for certain keywords. Replaced by broader terms from thesaurus.
2. Repeated words collected. Replaced with broader terms
3. Hierarchical network obtained
Experimental target

- Target word
  - Japanese Verb “osou”

- Target data derived from bilingual corpus (Utiyama and Isahara 2003), web texts, electronic novels, and electronic dictionary definitions.
  - Total data size 650 MB (approx.)

- Thesaurus that has hierarchical structure
  - NTT Nihongo Goi Taikei (hereafter “Taikei”)
Word replacements

- Used *Taikei* to replace words in extracted usages
- Word meanings reflected in semantic hierarchy where possible
  - Polysemy of words
    - Each broader terms replaced with word that included plural semantic attributes and was polysemous
  - Handling of words that are not in *Taikei*
    - Not handled in this study
Results

- Using a KWIC tool, extracted usages of “osou”
  - The string “(ni | ga) oso (wa | i | u | e | ttsu)”
    - “ni oso wa” (be hit | attacked | ... by)
    - “ga oso” (hit | attack ...)
  - 1103 usages extracted.

- Results of experiment described in next figure
  - Only items with two or more subordinate items shown.
  - Items surrounded by dashed frames are necessary for hierarchical network. Frames connected to other frames that do not have subordinate usage example expressed before are also shown.
Results (cont.)
Discussion

- Semantic frame obtained is very similar to Kuroda and Isahara’s.
- Found a semantic frame not in Kuroda's early version of analysis: “emotional and similar” attacks (F15).
- Method useful for semantic frame analysis during early stage of this kind of research because reduces effort to detect frames and frame elements of the target word.
New frame

- Event
  - Natural phenomenon
    - Vital phenomenon
      - Well-being / defective 83
        - Defective 83
          - Disease 81
    - Non-vital phenomenon
      - Meteorological phenomenon 126
        - Atmospheric phenomenon 126
          - Natural disaster 43
  - Human activity
    - Mental activity
      - Emotion 76
        - Sense 38
Discussion

- Semantic frame obtained is very similar to Kuroda and Isahara’s.

- Found a semantic frame not in Kuroda's early version of analysis: “emotional and similar” attacks (F15).

- Method useful for semantic frame analysis during early stage of this kind of research because reduces effort to detect frames and frame elements of the target word.
Summary

- Developed a method of constructing hierarchical semantic frame network, which is one goal of frame semantic analysis.
- Discovered the new semantic frames, and determined the relationship between discovered semantic frames using method.
- Our method can help analyze semantic frames more efficiently than by merely extracting usages from a corpus.
Thank you for your attention