Automatic extraction of topic hierarchies based on WordNet

Gerhard Brey[†] Miguel Vieira
Department of Digital Humanities, King's College London

Given by Jamie Norrish

Digital Humanities Australasia 2012

Overview

- Approach
- Corpus
- Workflow
  - Data preparation
  - Extraction of topics/target terms
  - Generation of topic tree
  - Examples
- Evaluation
  - Problems
  - Future work

Aim

Automatic generation of a topic hierarchy to be used for a faceted search interface

Castanet Algorithm

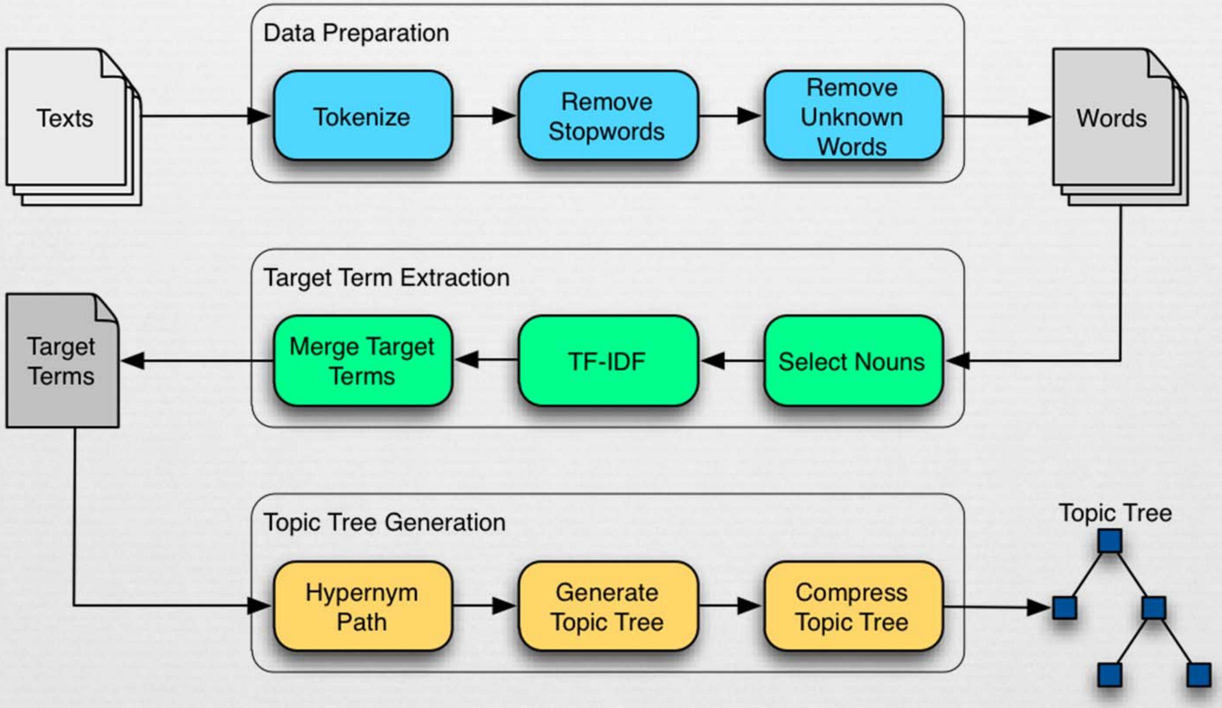
- Motivated by research described in 2 publications:
  - Emilia Stoica, Marti A. Hearst, Megan Richardson:
    "Automating Creation of Hierarchical Faceted Metadata Structures",
    Proceedings of NAACL/HLT 2007, Rochester, NY, April 2007, p. 244-251
    http://people.ischool.berkeley.edu/hearst/papers/castanet.pdf

  - Emilia Stoica, Marti A. Hearst:
    "Nearly-automated metadata hierarchy creation",
    Proceeding of HLT-NAACL 2004: Short Papers, Boston, Mass., 2004, p. 117-120
    http://www.aclweb.org/anthology/N/N07/N07-1031.pdf

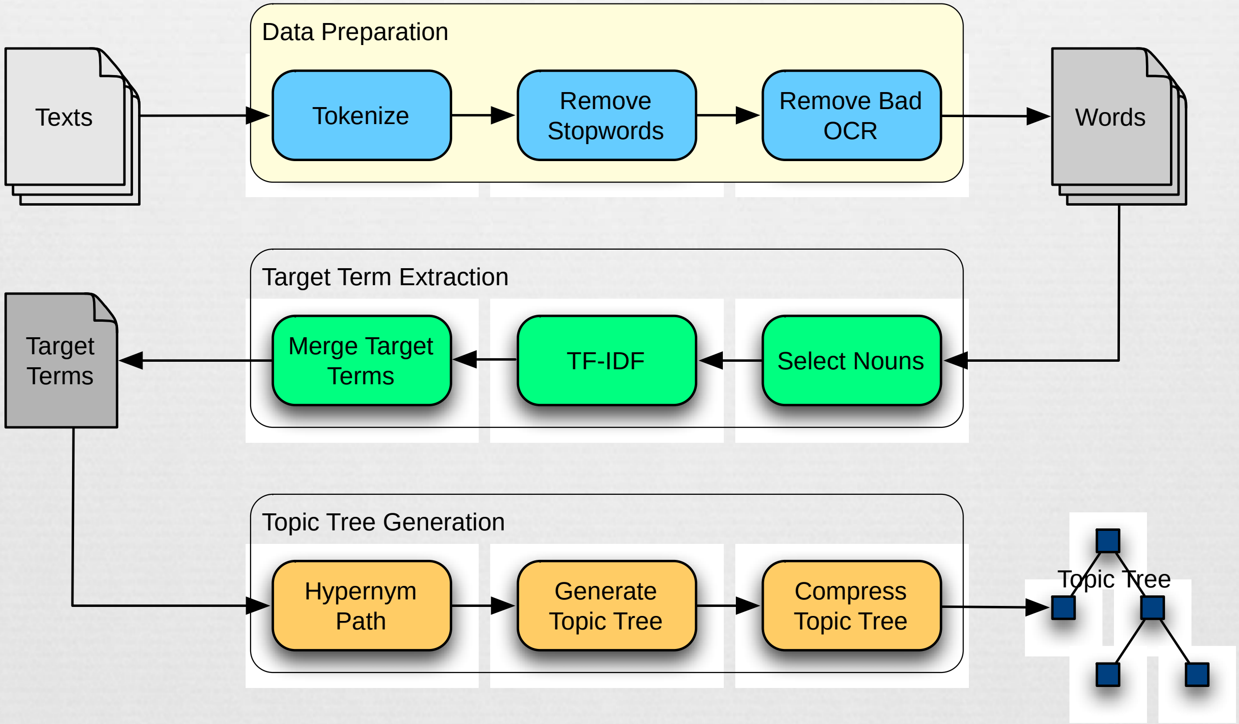
Corpus: English Women’s Journal (EWJ)

- Nineteenth Century Serials Edition (NCSE)
  - Digital edition of 6 newspapers and periodicals
- Edited and written by women
- Literary, political and social contents:
  - Education and employment of women
- One of the earliest newspapers committed to women’s rights in the nineteenth century
- Published between 1858 and 1864
  - 78 issues, 7964 articles

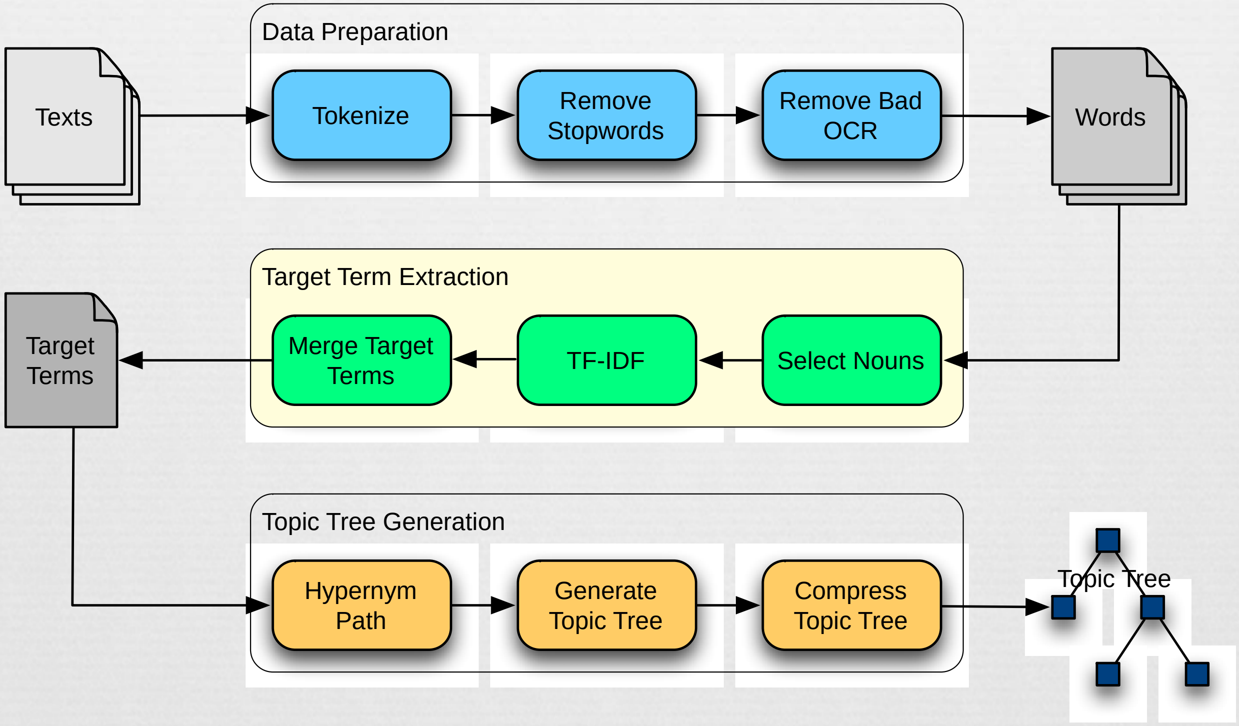
Workflow


Data Preparation
Texts
Tokenize
Remove Stopwords
Remove Unknown Words
Words
Target Term Extraction
Select Nouns
TF-IDF
Merge Target Terms
Target Terms
Topic Tree Generation
Hypernym Path
Generate Topic Tree
Compress Topic Tree
Topic Tree

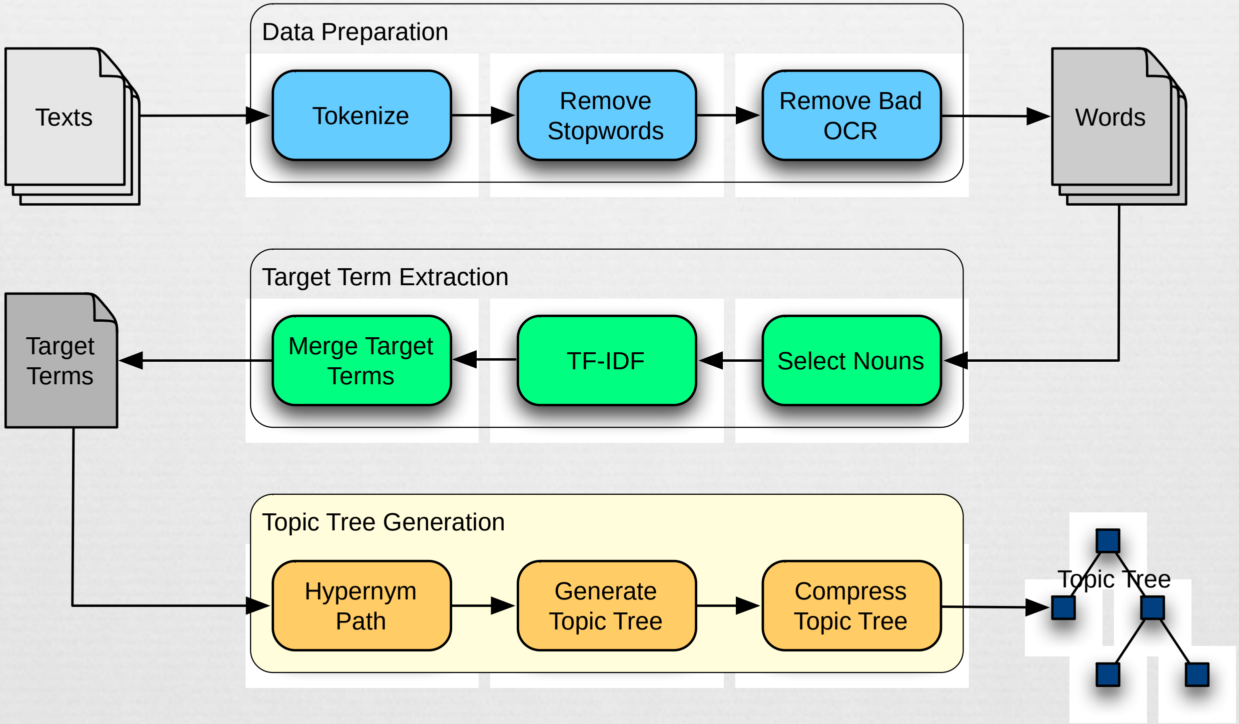
Data Preparation



Extraction of Target Terms



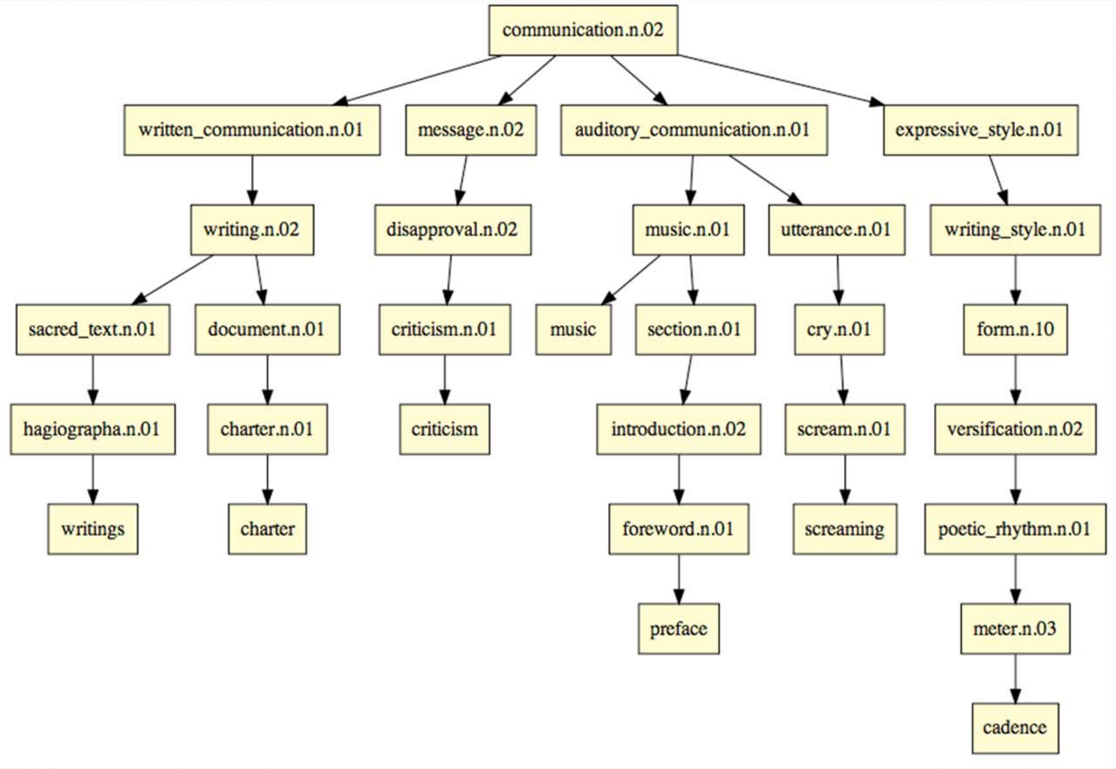
Generation of Topic Tree



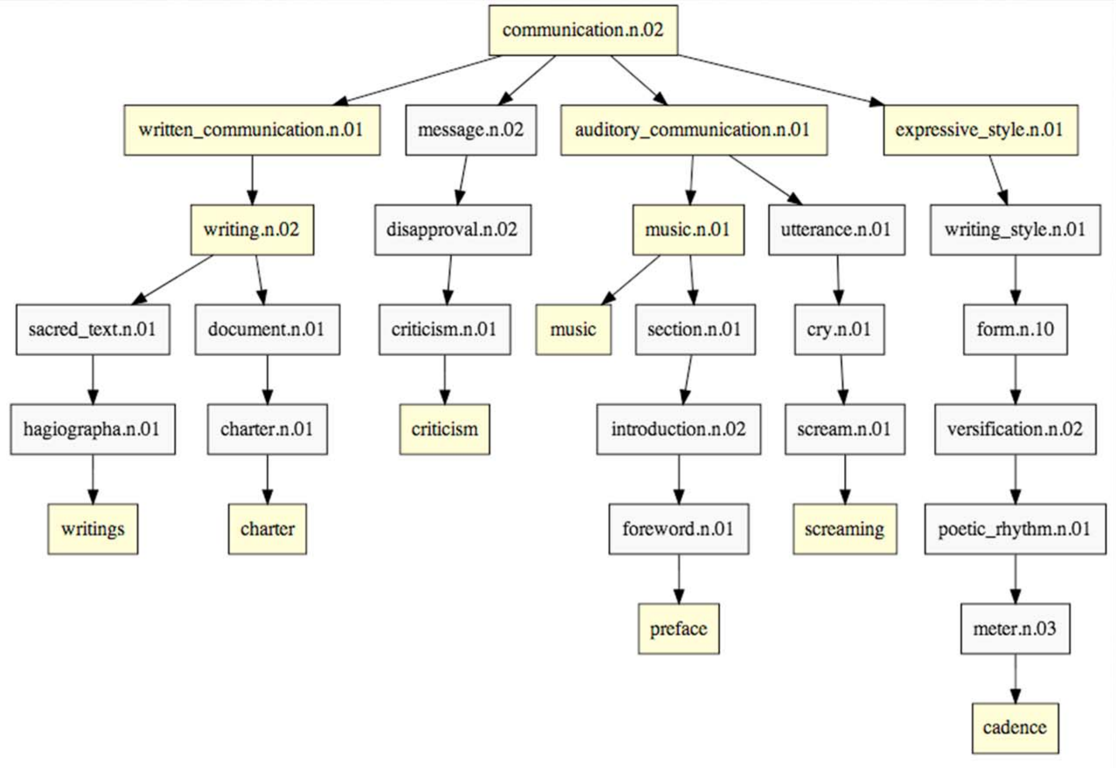
Hypernym Paths

- Without top level hypernyms “entity.n.01” and “abstraction.n.06”
- writings:
  - communication.n.02, written communication.n.01, writing.n.02, sacred text.n.01, hagiographa.n.01
- charter:
  - communication.n.02, written communication.n.01, writing.n.02, document.n.01, charter.n.01
- criticism:
  - communication.n.02, message.n.02, disapproval.n.02, criticism.n.01

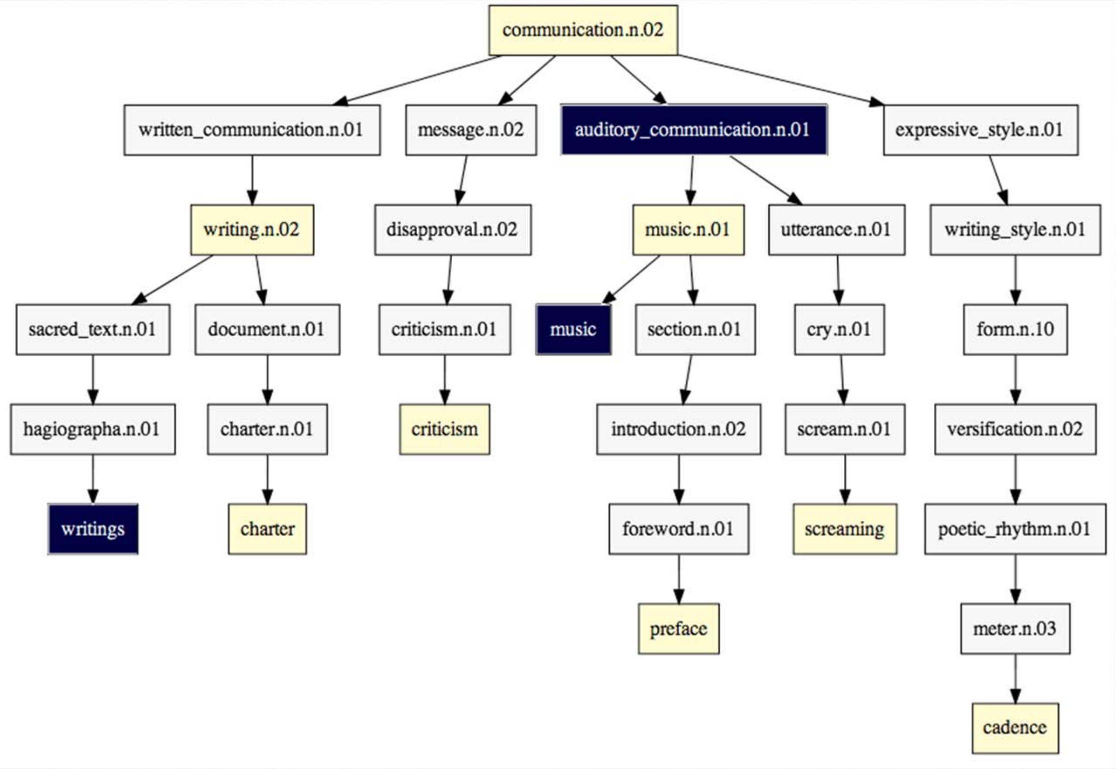
Tree Compression: before



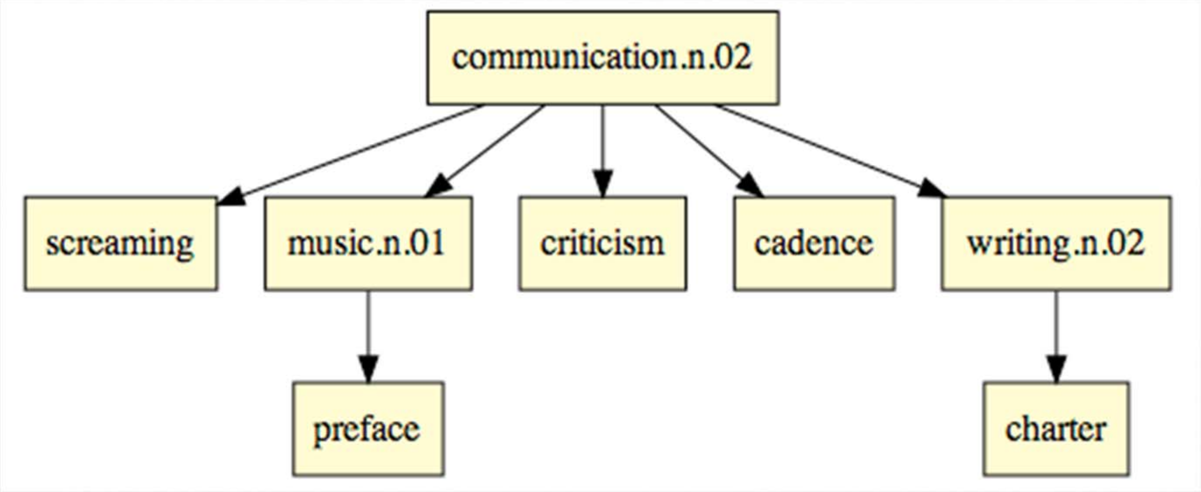
Tree Compression: removing parent nodes with less than 2 children



Tree Compression: removing leaves with similar names to parents



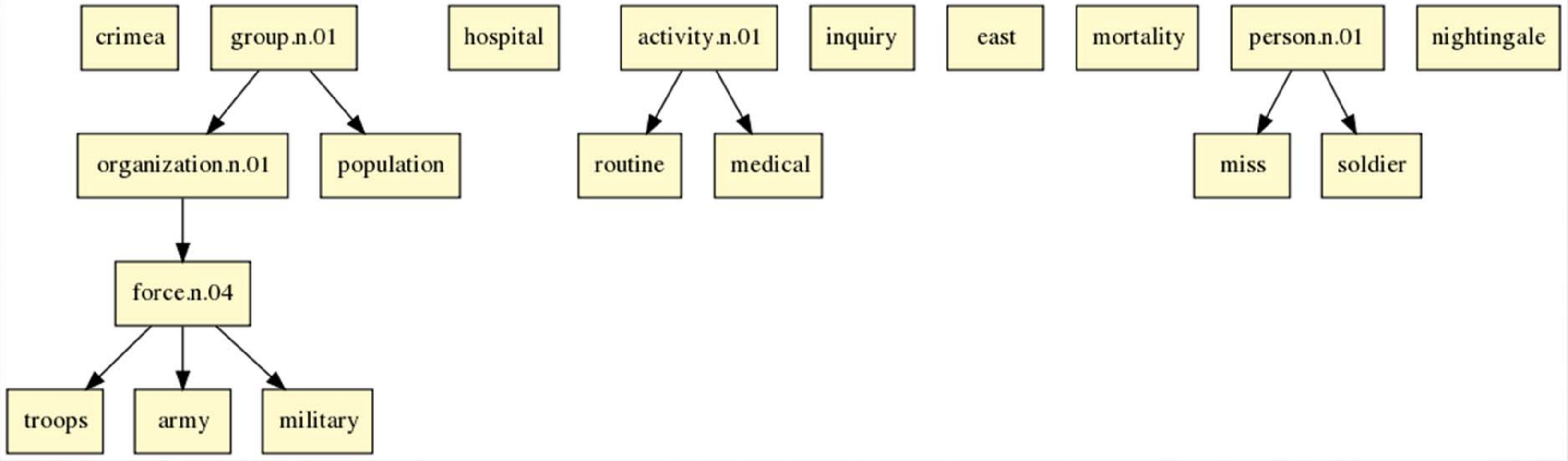
Tree Compression: after



Evaluation

- Test corpus
  - 1359 articles (minimum 300 characters)
  - 3.5 million words in total
  - 8013 unique target terms
  - Topic tree:
    - Before compression: 18234 nodes
    - After compression: 50% smaller
- Small subset evaluated, 20 articles
  - Generated topic hierarchy
  - Compared article content with topic hierarchy
  - Counted relevant terms for each article
- Over 90% of topics are relevant

Evaluation



Topic hierarchy for an article about Florence Nightingale and the Crimean war

Problems

- Misleading results because of mis-OCRed portions of text
  - Original: “Another act in the great European drama”
  - OCRed: “Ano , ther act in the great European drama”
  - Ano: Abu Nidal Organization
- Erroneous disambiguation of tokens with multiple meanings in WordNet
  - Drama, from above, added as a theatrical play

Future Work

- Evaluation of output by domain experts
- Explore other algorithms to extract meaningful target terms
- Better handling of OCR errors
- Improvement of automatic disambiguation

Conclusion

- Resulting topic hierarchies are intuitive
- Topic hierarchies form an appropriate basis for a faceted search interface
- It can be applied to other corpora in other languages
  - Provided WordNet is available

Questions?

Miguel Vieira jose.m.vieira@kcl.ac.uk