Semi-automatic Construction of Cross-period Thesaurus

Chaya Liebeskind, Ido Dagan, Jonathan Schler
Bar Ilan University

Research objective

Support search over historical texts of two types:

- Ancient texts
- Modern texts quoting ancient language (mixed)

modern | car → wagon, carriage | ancient

Researchers are typically aware of modern language

Cross-period (Diachronic) Thesaurus

Target term (modern)	Ancient related terms	Modern related terms
Car	wagon, carriage	auto, motor
Ship	Dreadnought, Bireme	boat, craft
Weapon	sword, naboot	gun, bomb

Our task:

- Support searches of cultural resources
- High-quality diachronic thesaurus construction
  - ➢ Semi-automatic setting

Hardly explored computationally

Automatic generation of candidates

Statistical approaches for semantic relatedness identification

- First-order co-occurrence (fits corpus nature)
- Second-order distributional similarity
  - ☹ mediocre quality (corpus dependent)

Co-occurrence-based methods

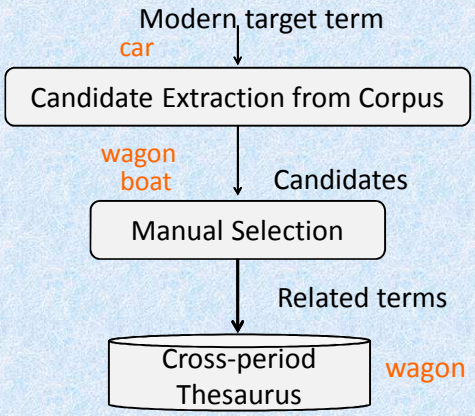
Assumption

Words that occur frequently together are topically related

Common metrics

Dice coefficient, Pointwise Mutual Information, log-likelihood

Semi-automatic thesaurus construction (non-iterative)



Limited recall

☹ Only documents consisting the modern target term are utilized (only mixed corpora, not ancient)

Solution

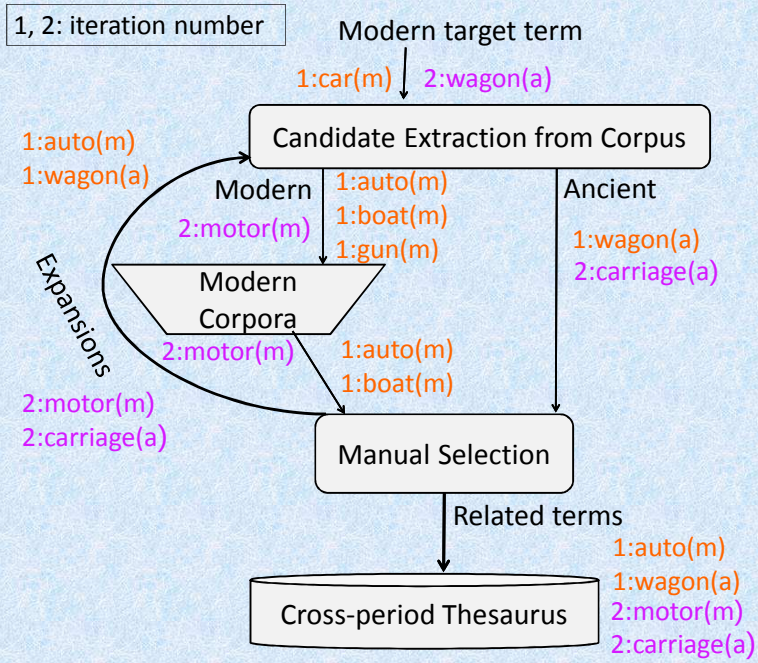
Increase number of documents in the statistical extraction

- Utilize ancient corpora

Apply query expansion to increase coverge

car → auto → wheels

Iterative semi-automatic scheme



- Candidates are Multi Word Expressions (MWEs)
  - ➢ Generated by statistical association metrics *Mutual Expectation*, *Mutual Rank Ratio*
- Period classification - by corpus type

Case study: Cross-period Jewish thesaurus

The Responsa Corpus

- Questions posed to rabbis along their detailed answers
- Written over about a thousand years
- 76,760 articles
- Used for previous IR and NLP research

Ancient Corpora

➢ Responsa documents from the 11th century until the 19th century

Mixed Corpora

➢ Responsa documents from the 20th century until today
➢ Jewish law literature

Results

Evaluation: over 100 modern test target terms

Compare the iterative scheme to a baseline of a similar non-iterative method

- Compare for same number of judgements

Method	Baseline	Iterative	Rel. Δ
# judgments (J)	14626	14462	–
# related terms (RT)	892	1106	0.24
Average RT per target term	9	11	0.24
Precision-Productivity (RT/J)	0.06	0.08	0.33
Relative Recall (R)	0.78	0.99	0.27

Conclusions and future work

- A new task: cross-period thesaurus construction
- Semi-automatic iterative QE scheme
- Increasing thesaurus coverage, while optimizing the lexicographer manual effort

Future work:

- Adopt second-order distributional similarity methods