Continuous Access To Cultural Heritage

a computer science research programme

Enriching a Thesaurus to Improve Retrieval of Audiovisual Documents

Laura Hollink, *Véronique Malaisé* and Guus Schreiber

Background

- Research conducted within MUNCH and CHOICE@CATCH
- Concrete use case: the Dutch Institute for Sound & Vision
- Munch's research domain: multimodal search in AV archives
- Choice's research domain: extraction and ranking of keywords for semi-automatic annotation
- Meeting point: the thesaurus
  - used for query expansion
  - used for keywords extraction and ranking
  - Need of structure!

→ does the inference of relationships from external resources help?

Outline

- The experiment's background:
  - searching in manually annotated AV archives
  - query expansion
- The thesaurus: content, enrichment and anchoring
- The experiment: Does adding structure derived from an external resource improve the thesaurus based search in AV archives?
  - audiovisual data: subset of the archives submitted to TRECVid
  - queries and ground truth from TRECVid

Searching in large manually annotated AV archives

- Searching in the metadata (no image-based search at S&V):
  - full text search in the document's description
  - keywords search in metadata fields with constrained value
- Annotation process: few keywords assigned per document, as specialized as possible
- Time consuming operation: 4 hours of work for the full documentation of one hour of program
- Description of the main topic(s) of a document only

  → very good precision, possibly low recall

Query expansion

- Principle: "Query expansion is the process of adding additional terms to the original query in order to improve retrieval performance"[1]
- Option 1 (lexical level): adding synonyms [2]
- Option 2 (structural level): adding terms at different levels of specialization in the thesaurus [3]
- The thesaurus has to have a rich structure for the second option to be successful!

Thesaurus structure

- Hierarchical relationships: BroaderTerm/NarrowerTerm
- Associative relationships: RelatedTerm
- Linguistic relationships: Use/UsedFor
- Sometime facets (Subject, People, Places,...)
- No constraint to have one single root
- Multiple hierarchies are allowed
- Lack of formal semantics, meant for human interpretation
- In real-case applications, the structure is often not very rich

Possible solution

- Adding new relationships from external resources
- BUT: it might jeopardize the precision, along with (not?) increasing the recall
- Experiment on real data from the Netherlands Institute of Sound and Vision

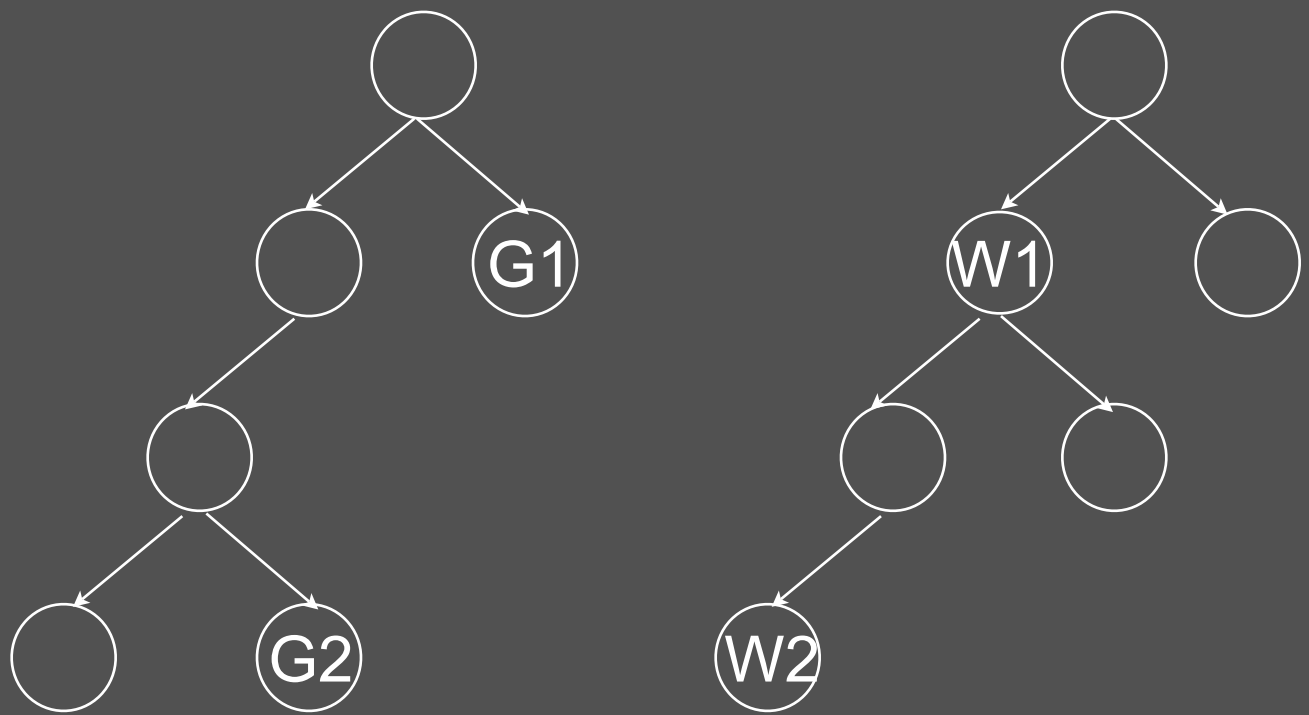
The thesaurus of Sound and Vision

- Faceted Thesaurus 160.000 terms:
  - ~3800 Subject keywords (~2000 NonPreferred Terms),
  - ~97.000 "Person"s,
  - ~27.000 "Names",
  - ~14.000 Locations,
  - 113 Genres and
  - ~18.000 Makers.
- Subject facet: BroaderTerm, NarrowerTerm, RelatedTerm, Use/Use For
- Up to 6 levels of depth in hierarchy but 3708 Terms between 1-3 levels of depth
- SKOS representation

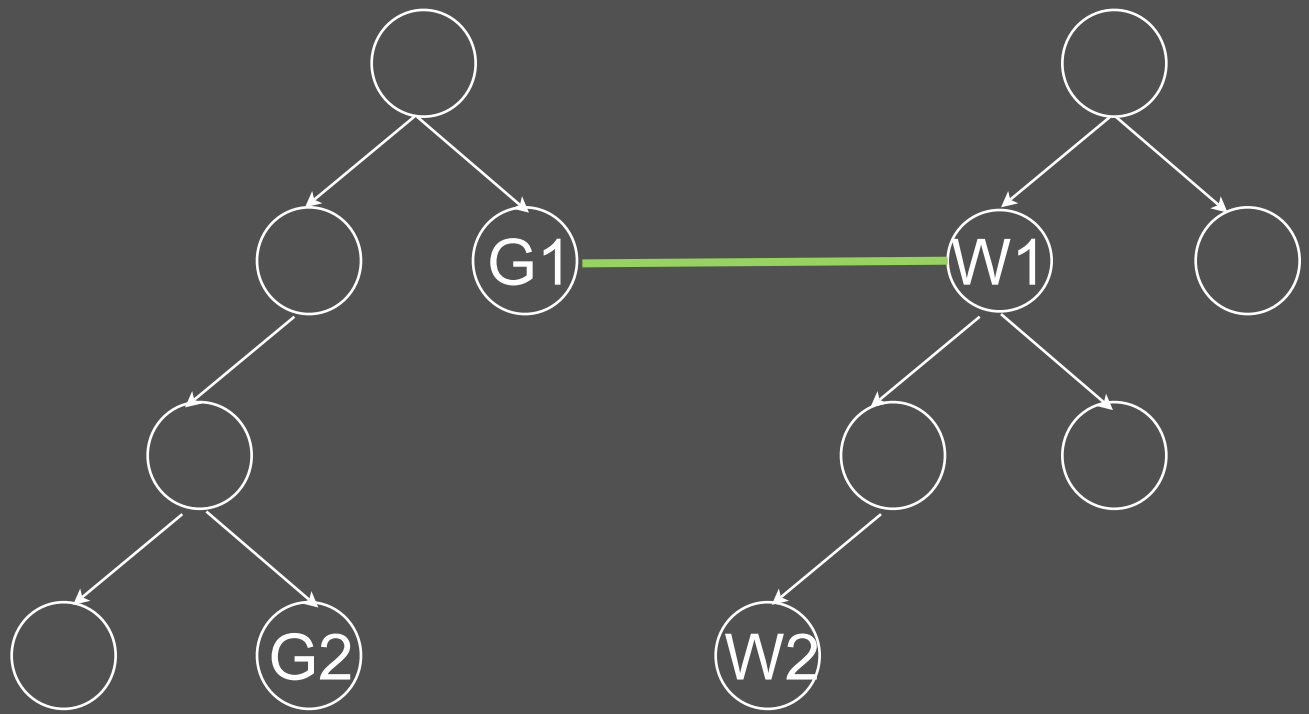
WordNet

- Terminological database created for the English language and maintained at the Princeton University
- 155,287 English words: nouns, verbs, adjectives and adverbs
  - words are divided by meanings: 3 different “Tree” (woody plant, graph and actor’s name) are in 3 different synsets
  - words are grouped by meaning: 3 different words for *Cliff* (cliff, drop and drop-off) are in the same synset
  - Synsets are organized in relationships and described by a Gloss
- RDF/OWL version of WordNet 2.0 released by W3C: http://www.w3c.org/TR/wordnet-rdf/

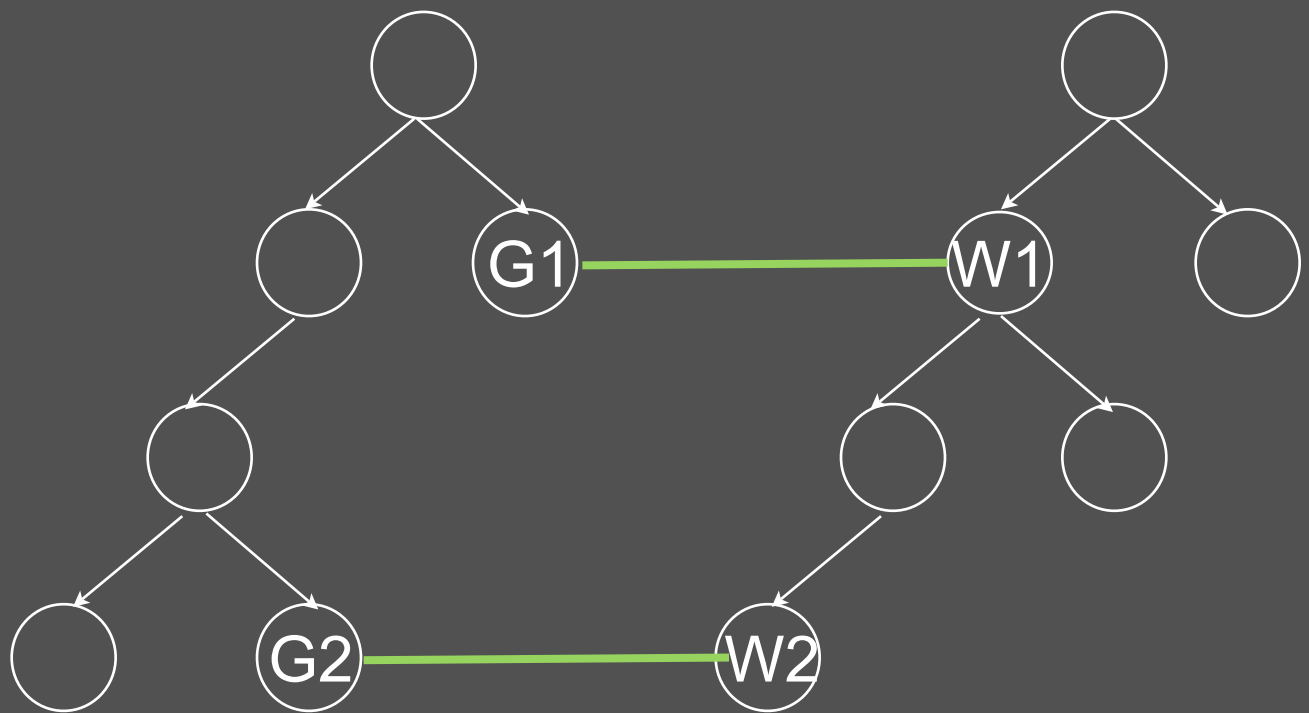
Thesaurus enrichment: adding relationships by anchoring to WordNet



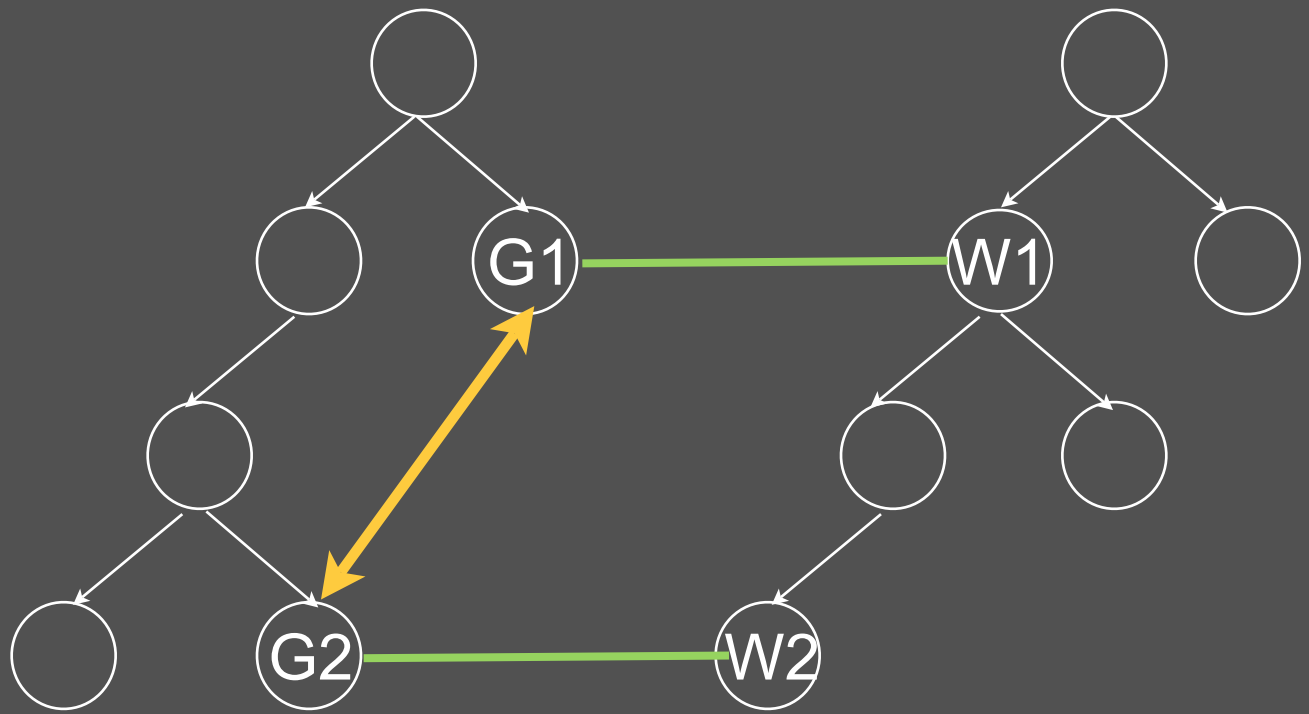
Thesaurus enrichment: adding relationships by anchoring to WordNet



Thesaurus enrichment: adding relationships by anchoring to WordNet



Thesaurus enrichment: adding relationships by anchoring to WordNet



The anchoring of GTAA (Dutch) to the English WordNet

- Anchoring: link from one resource to another, for mapping for example
- Option 1: by comparing labels
- Option 2: by comparing descriptions

The anchoring process

- 1/ Enriched the GTAA on the lexical level, to get the best possible coverage
  - added synonyms from online dictionaries,
  - got singular and
  - decomposition of complex/multiword terms from Celex
- 2/ Queried an online bilingual dictionary, to get terms description in English
- 3/ Compared the definitions with WordNet glosses: 99% are exact matches!

→ 1855 Subject terms from GTAA were anchored in WordNet

Anchoring results

- Previous experiment:
  - 2222 GTAA terms matched the online dictionary
  - these represent 1748 unique entries
  - keeping only Nouns and Verbs -> 1655 entries and 7530 def
  - mapping to 1060 Synsets (1 term to multiple synsets, some synsets to multiple terms)
- new version: 1855 synsets, to be fully evaluated
  - more relationships expected than the exact match intended in the previous experiment

The experiment: objectives

- Measuring the retrieval performances of the newly inferred relationships as **compared** with the original ones
- Measuring the retrieval performances of the newly inferred relationships as **combined** with the original ones

The experiment's setting

- Nine runs of query expansion (exact matches, as baseline) based on:
  - broader, narrower, related terms to GTAA terms, and the full combination
  - distance 1,2,3 in WordNet and the combination of these
  - all of the above relationships

The dataset

- TRECVid 2007 dataset for high-level feature extraction task
  - 50 hours of news magazines, science news, news reports, documentaries, educational programs and archival material from Sound and Vision
  - 36 queries (features)
    - manually translated to GTAA, 3 had no translation
    - 9 queries too generic (in 2/3rds of programs)
  - a manually constructed ground truth (relevant, irrelevant, not checked)
    - adapted it to be on the document level

→ 104 TV programs, 3.6 GTAA per program, 25 queries and ground truth of averagely 27 doc per query

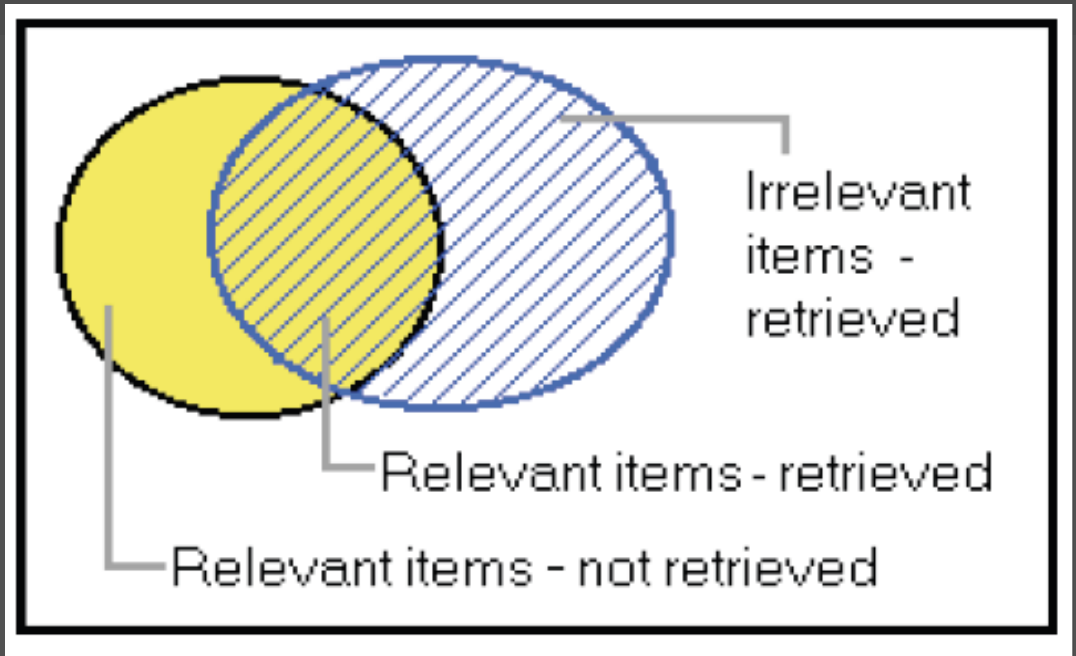
Evaluation

- For each run:
  - Precision
  - Recall

  - Harmonic mean: the F1-measure
    - $F_1 = 2x((Prec \times Rec)/(Prec + Rec))$

Evaluation

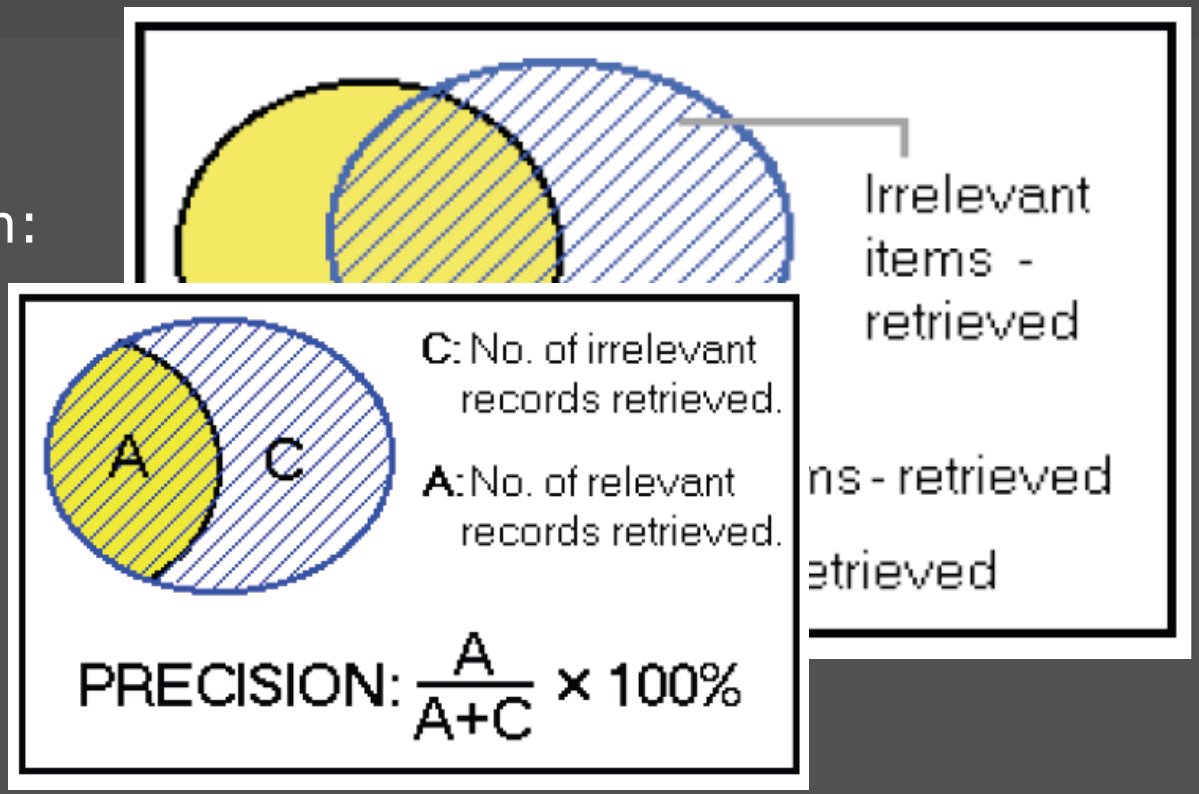
- For each run:
  - Precision
  - Recall



  - Harmonic mean: the F1-measure
    - $F_1 = 2x((Prec \times Rec)/(Prec + Rec))$

Evaluation

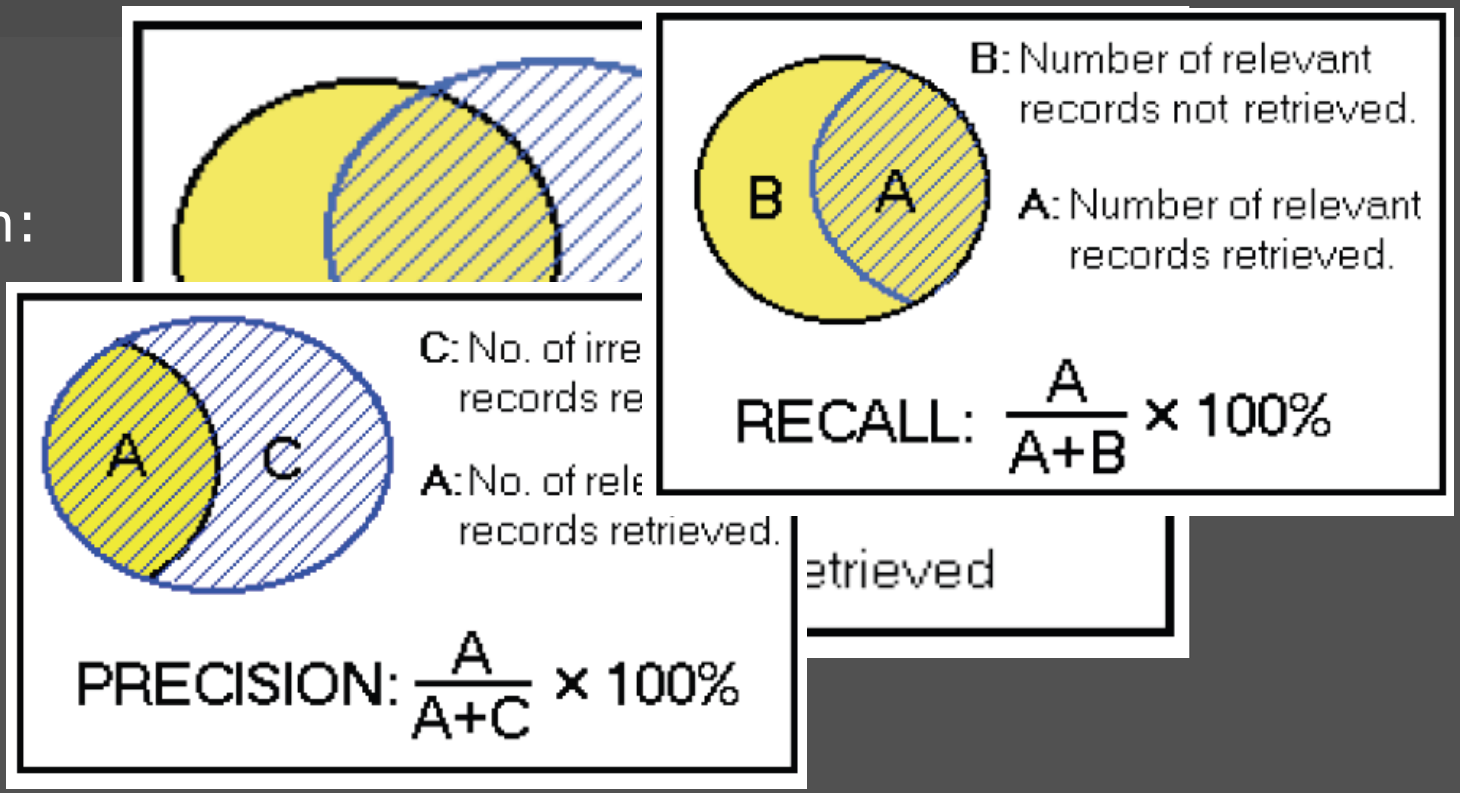
- For each run:
  - Precision
  - Recall



  - Harmonic mean: the F1-measure
    - $F_1 = 2x((Prec \times Rec)/(Prec + Rec))$

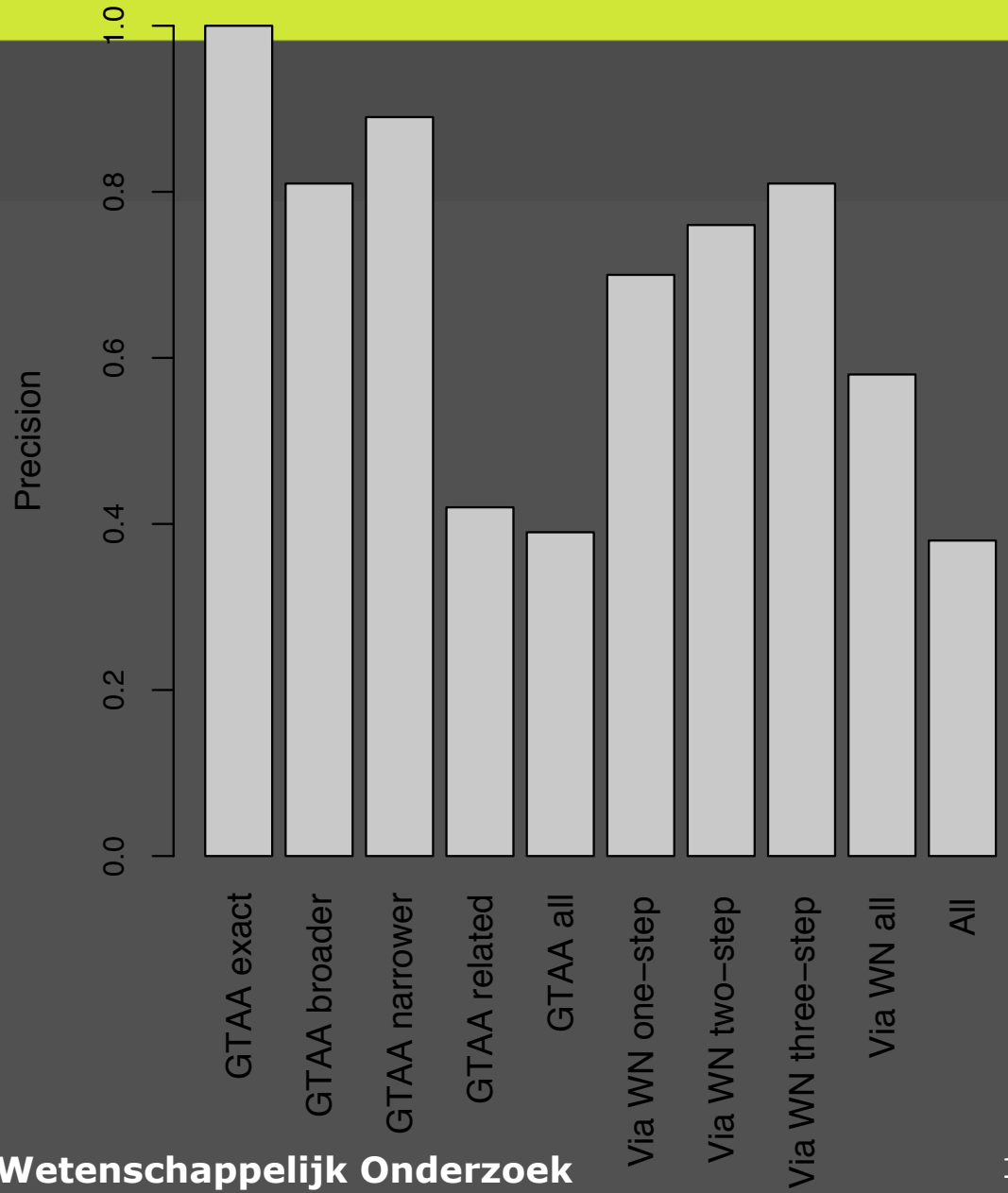
Evaluation

- For each run:
  - Precision
  - Recall

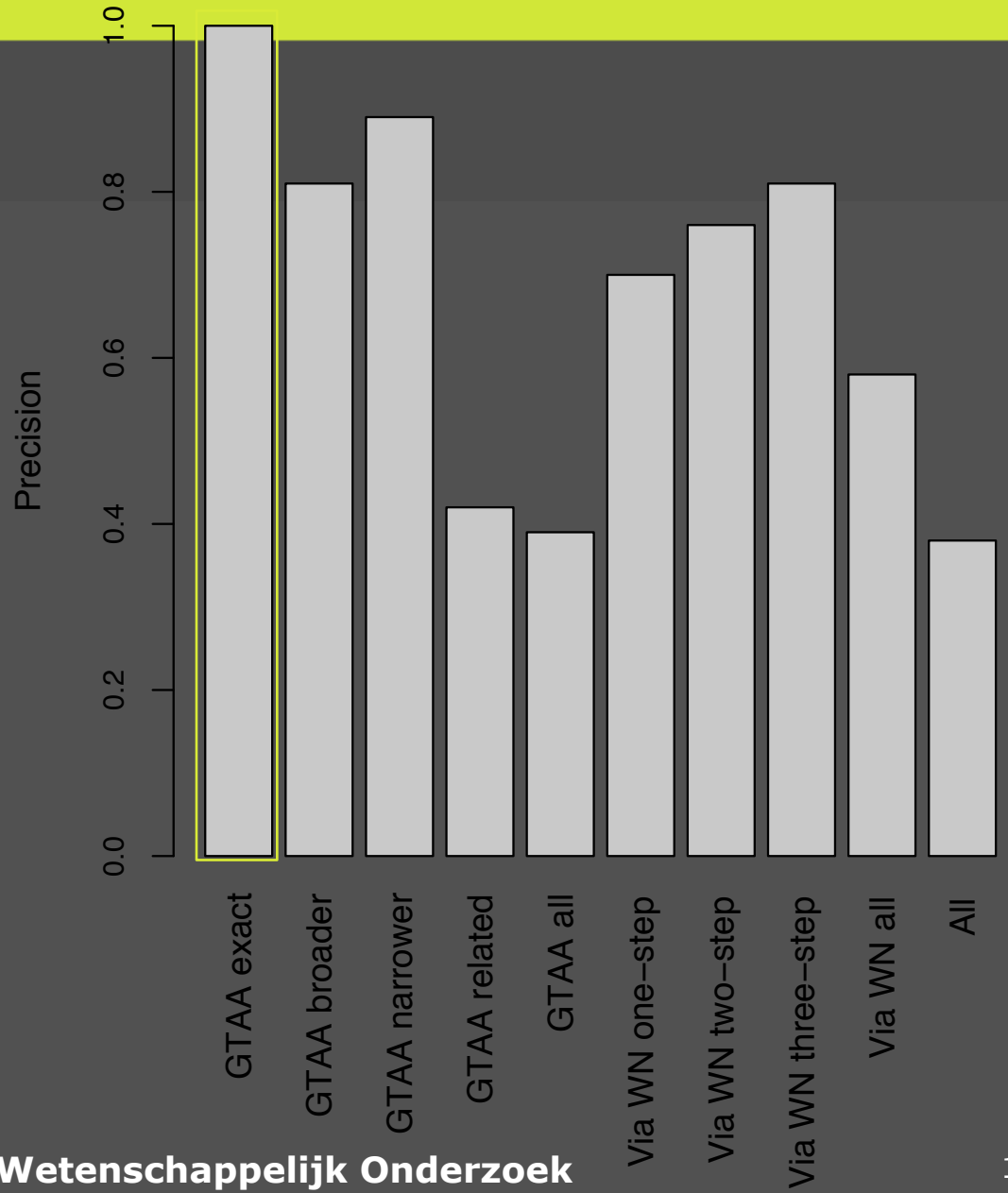


  - Harmonic mean: the F1-measure
    - F_1= 2x((Prec x Rec)/(Prec + Rec))

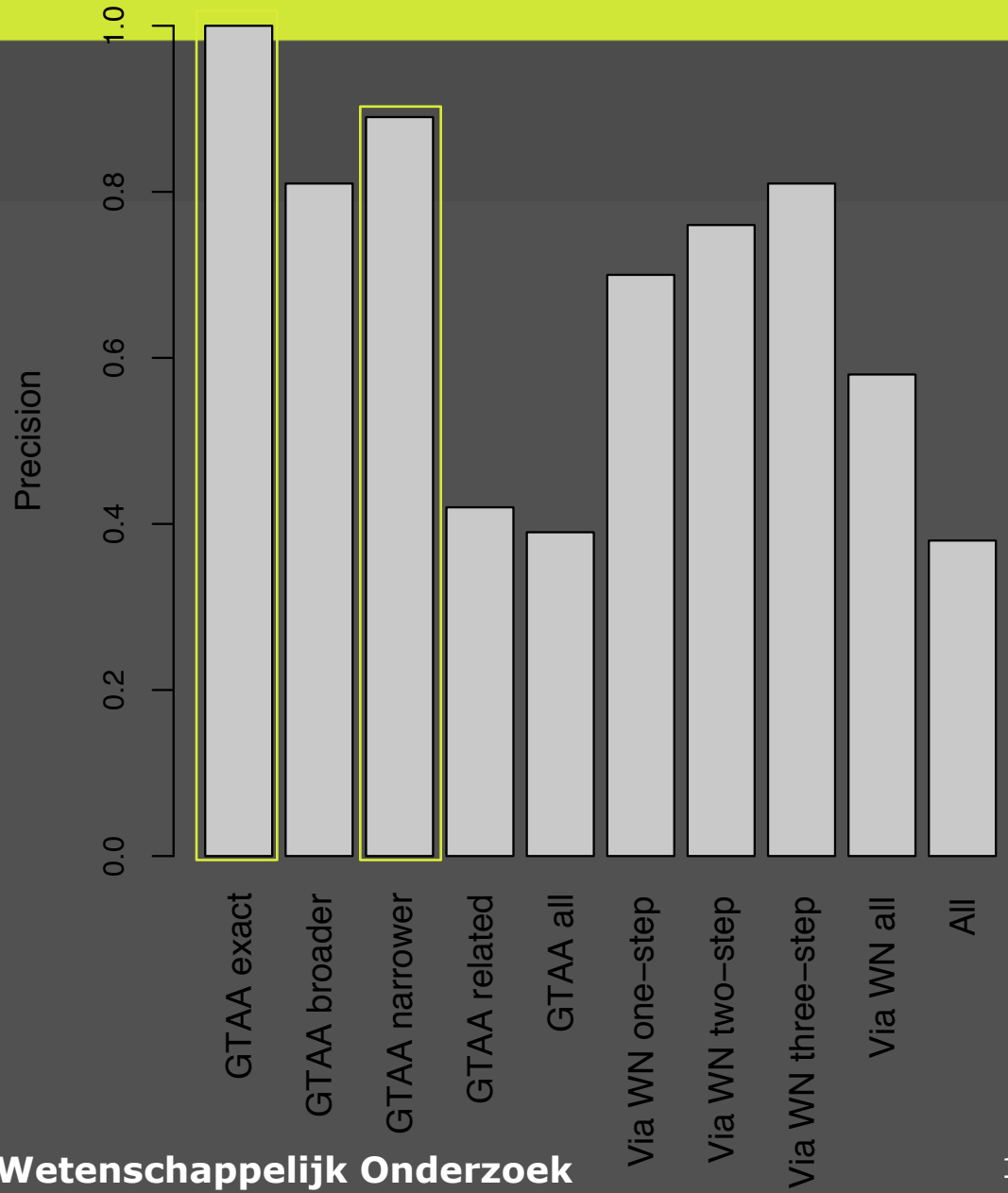
Results


Precision
0.0
0.2
0.4
0.6
0.8
1.0
GTAA exact
GTAA broader
GTAA narrower
GTAA related
GTAA all
Via WN one−step
Via WN two−step
Via WN three−step
Via WN all
All

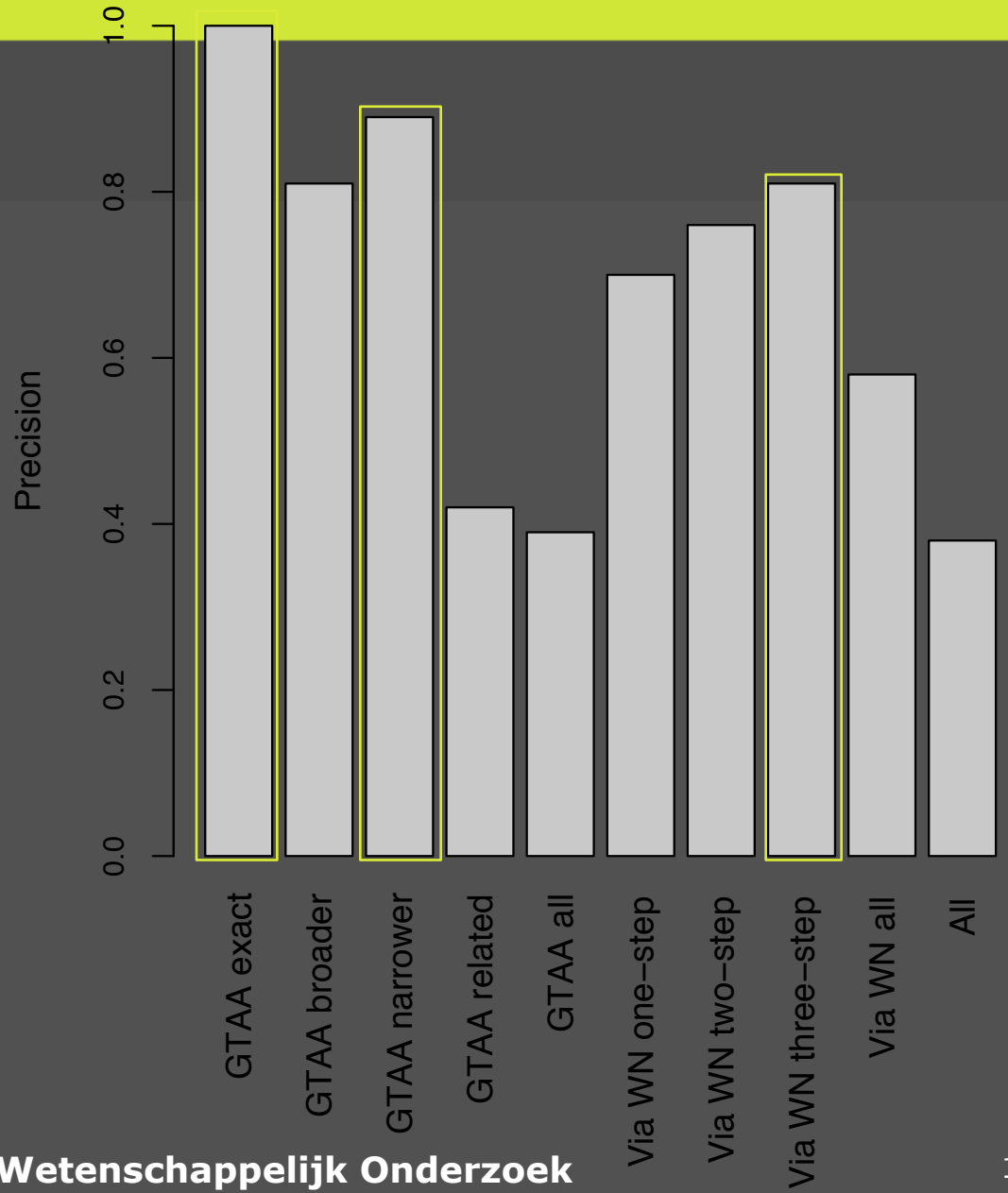
Results


Precision
0.0
0.2
0.4
0.6
0.8
1.0
GTAA exact
GTAA broader
GTAA narrower
GTAA related
GTAA all
Via WN one−step
Via WN two−step
Via WN three−step
Via WN all
All

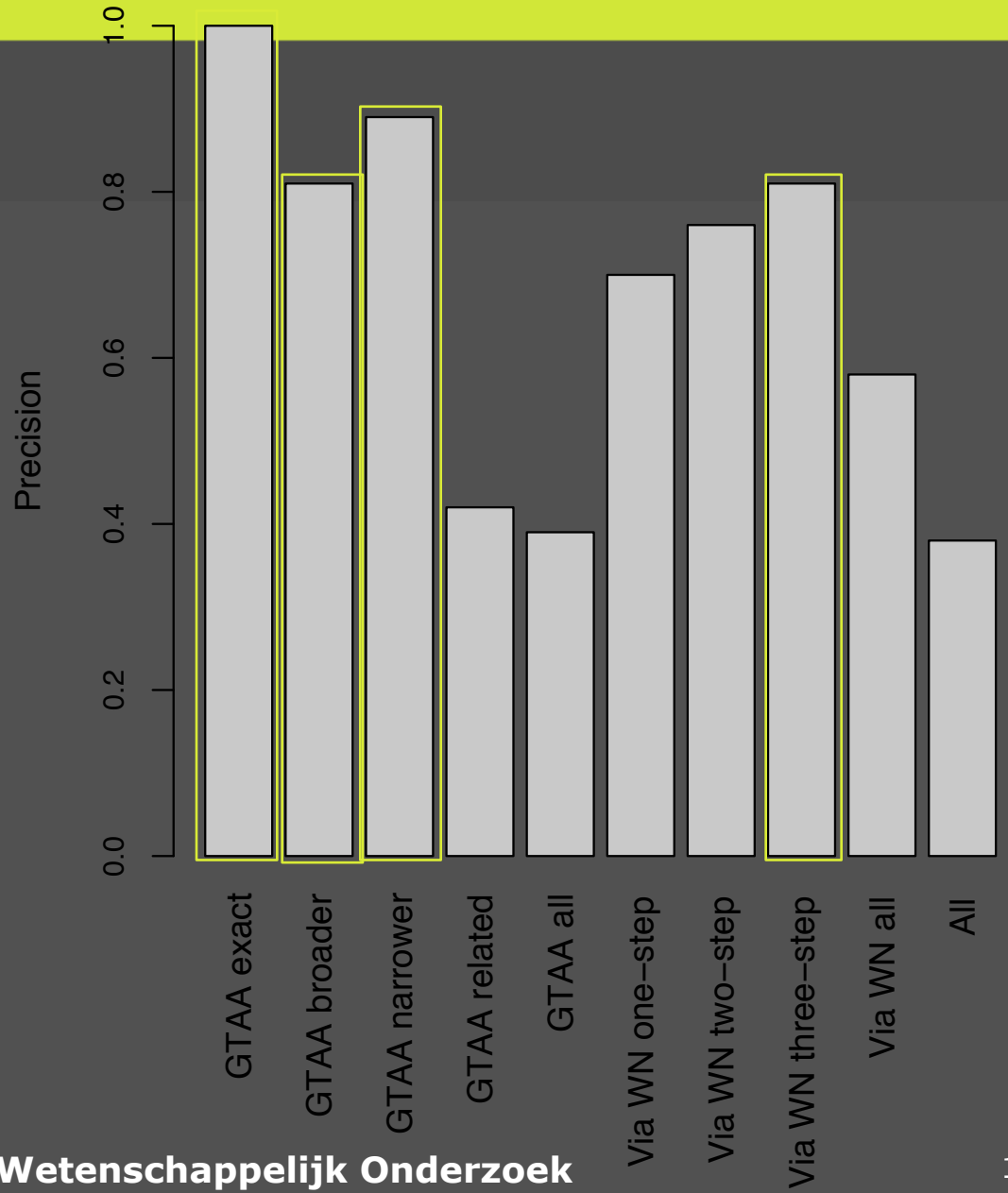
Results

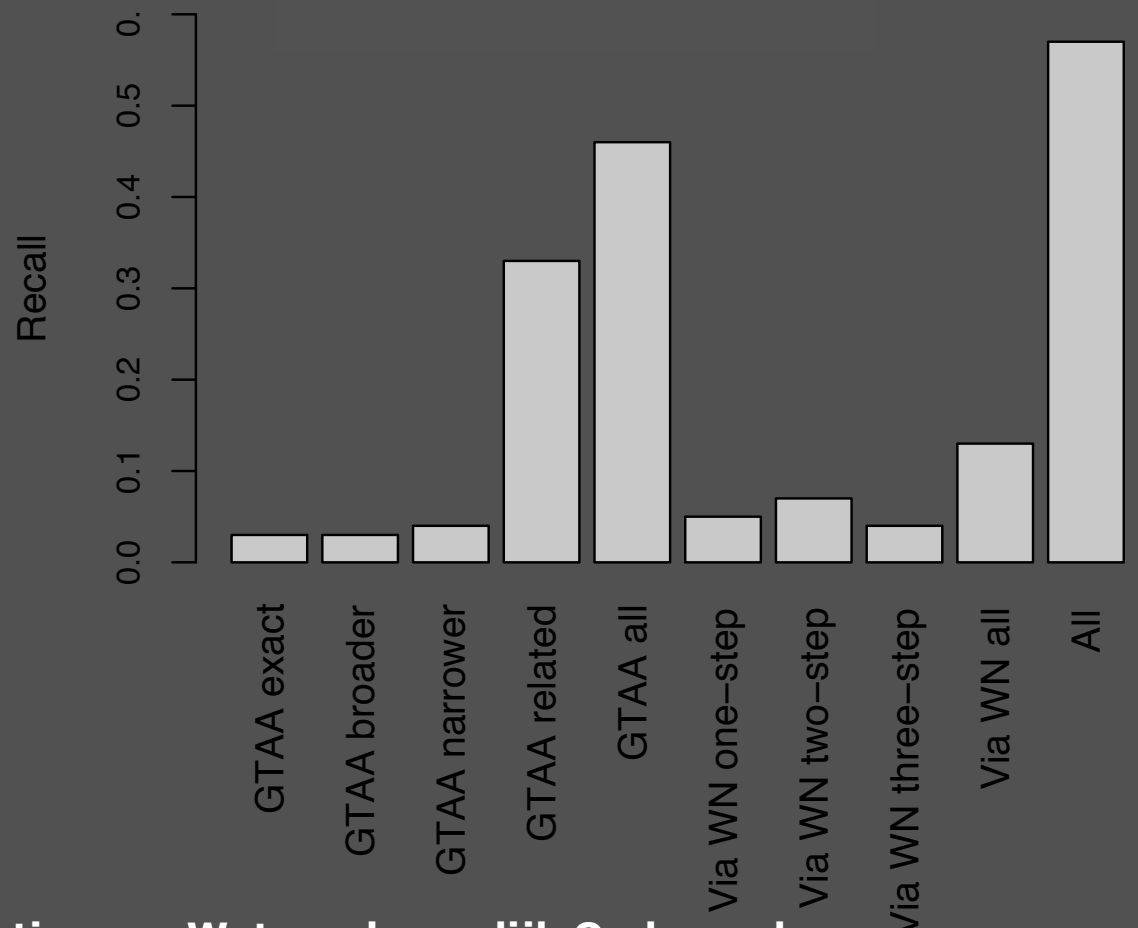

Precision
0.0
0.2
0.4
0.6
0.8
1.0
GTAA exact
GTAA broader
GTAA narrower
GTAA related
GTAA all
Via WN one−step
Via WN two−step
Via WN three−step
Via WN all
All

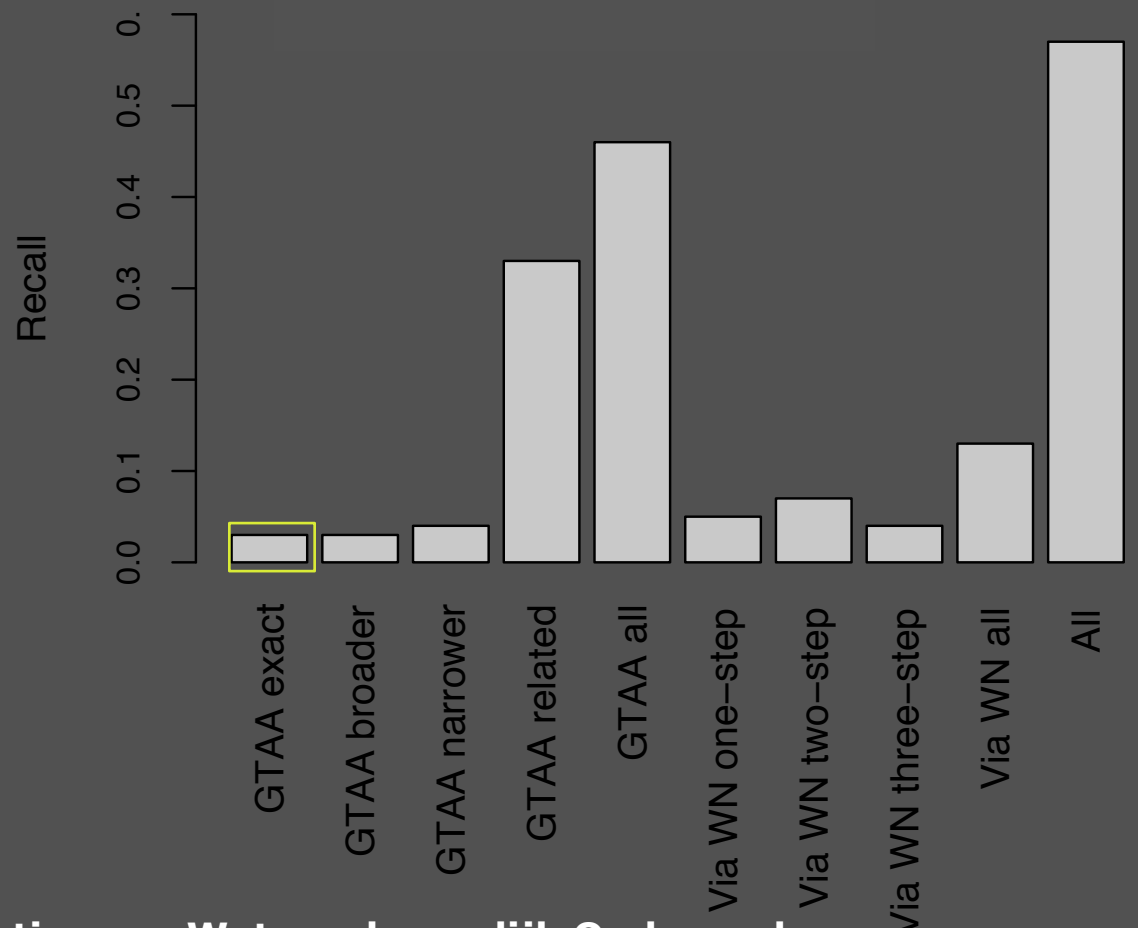
Results

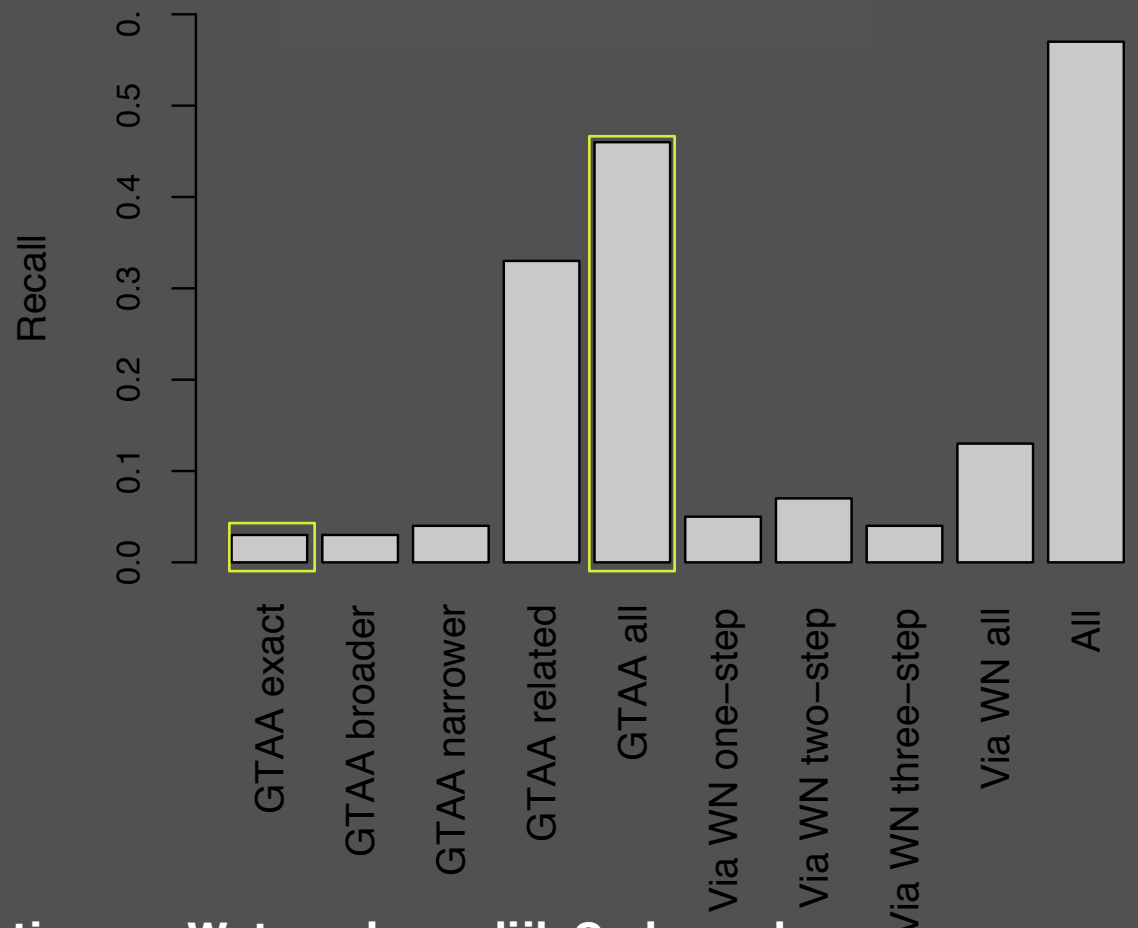

Precision
0.0
0.2
0.4
0.6
0.8
1.0
GTAA exact
GTAA broader
GTAA narrower
GTAA related
GTAA all
Via WN one−step
Via WN two−step
Via WN three−step
Via WN all
All

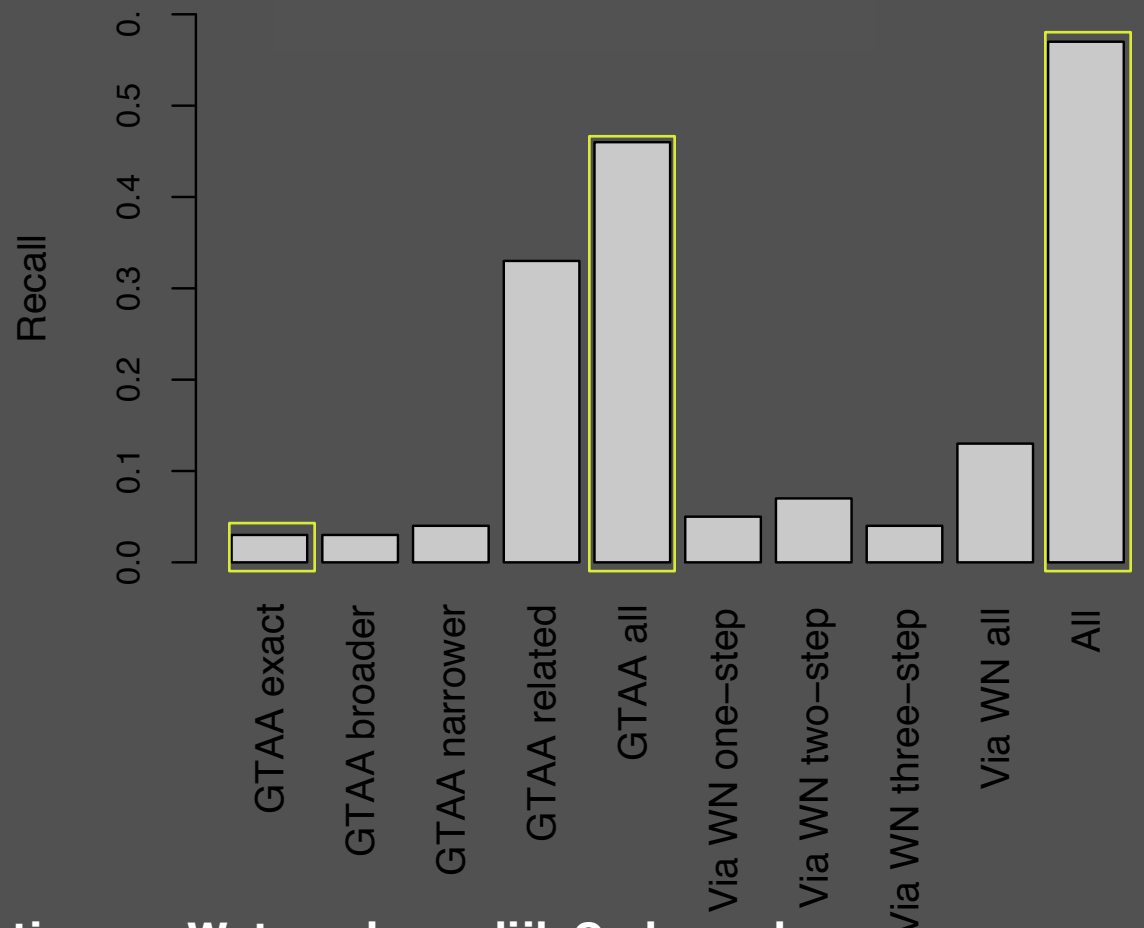
Results


Precision
0.0
0.2
0.4
0.6
0.8
1.0
GTAA exact
GTAA broader
GTAA narrower
GTAA related
GTAA all
Via WN one−step
Via WN two−step
Via WN three−step
Via WN all
All


Recall
0.0
0.1
0.2
0.3
0.4
0.5
0.
GTAA exact
GTAA broader
GTAA narrower
GTAA related
GTAA all
Via WN one−step
Via WN two−step
Via WN three−step
Via WN all
All

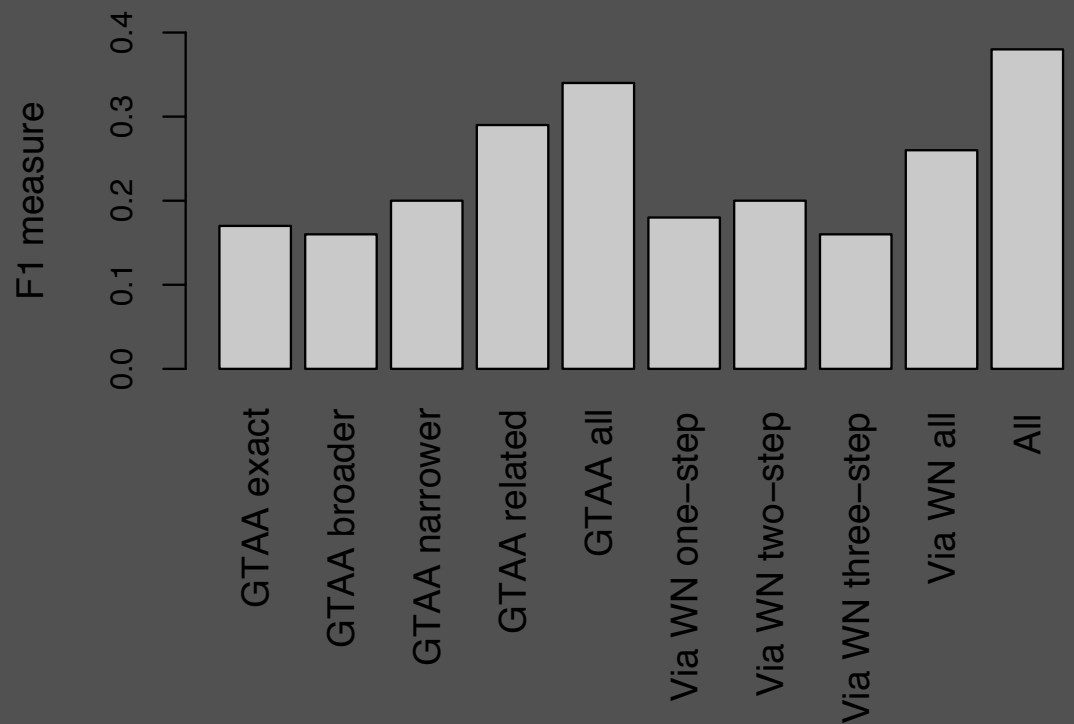

Recall
0.0
0.1
0.2
0.3
0.4
0.5
0.
GTAA exact
GTAA broader
GTAA narrower
GTAA related
GTAA all
Via WN one−step
Via WN two−step
Via WN three−step
Via WN all
All

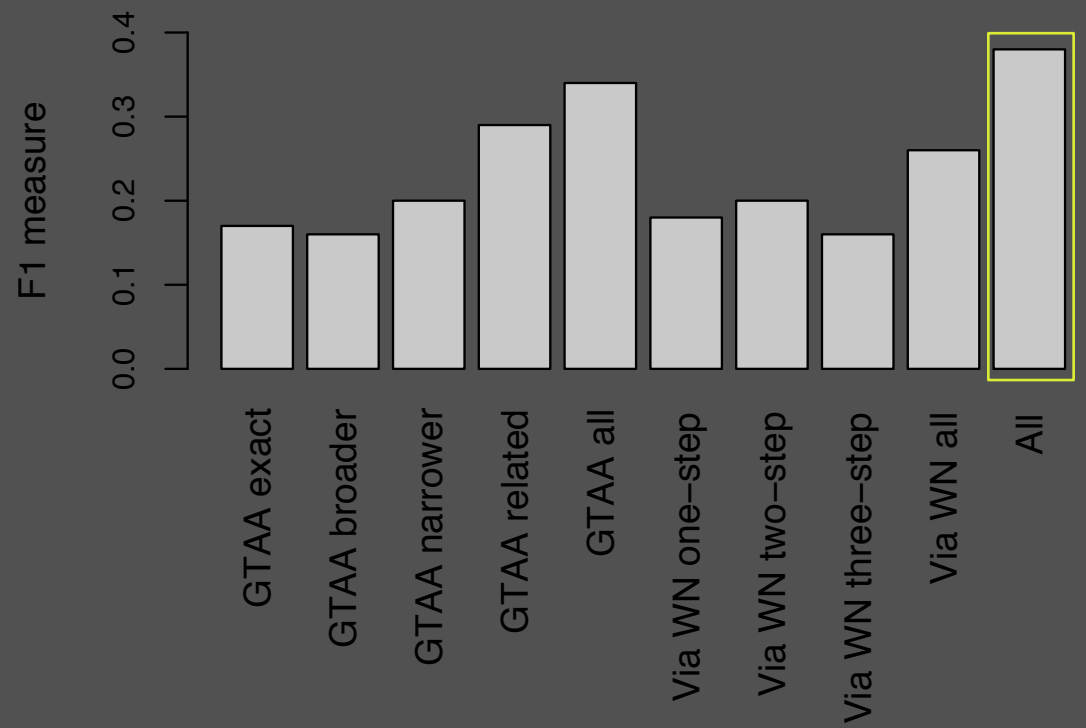

Recall
0.0
0.1
0.2
0.3
0.4
0.5
0.
GTAA exact
GTAA broader
GTAA narrower
GTAA related
GTAA all
Via WN one−step
Via WN two−step
Via WN three−step
Via WN all
All

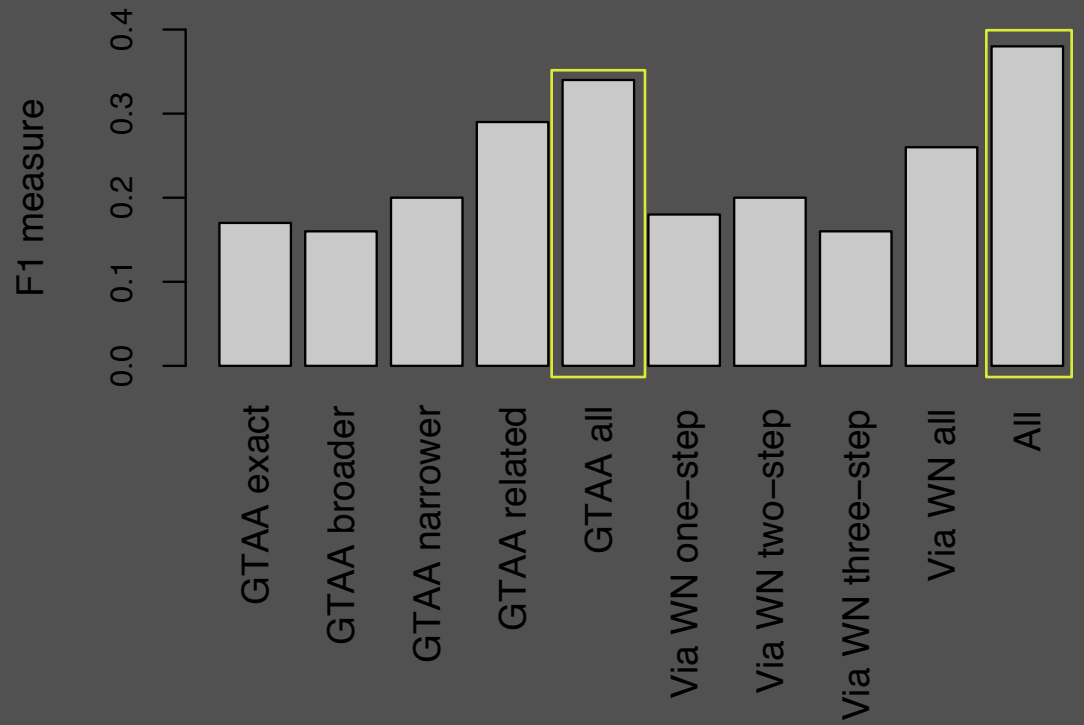

Recall
0.0
0.1
0.2
0.3
0.4
0.5
0.
GTAA exact
GTAA broader
GTAA narrower
GTAA related
GTAA all
Via WN one−step
Via WN two−step
Via WN three−step
Via WN all
All

Note about the results

– The evaluation is based on the manual evaluation of the TRECVid competition: only the shots that were returned by at least one participant were evaluated.


F1 measure
0.0
0.1
0.2
0.3
0.4
GTAA exact
GTAA broader
GTAA narrower
GTAA related
GTAA all
Via WN one−step
Via WN two−step
Via WN three−step
Via WN all
All


F1 measure
0.0
0.1
0.2
0.3
0.4
GTAA exact
GTAA broader
GTAA narrower
GTAA related
GTAA all
Via WN one−step
Via WN two−step
Via WN three−step
Via WN all
All


F1 measure
0.0
0.1
0.2
0.3
0.4
GTAA exact
GTAA broader
GTAA narrower
GTAA related
GTAA all
Via WN one−step
Via WN two−step
Via WN three−step
Via WN all
All

Conclusion

- The newly inferred relationships, with limited clean-up, enables to increase moderately the recall of a search task (at a comparable precision), over results obtained only with the help of existing relationships
- Using the new relationships only gives a f-measure that is comparable to the one obtained when using the "narrower term" relationship: they enable to get useful results when no structure is present
- The nature of the thesaurus relationship taken into account in this experiment does not seem to matter, at the difference with experiments where patterns were designed to improve the results

Future work

- Evaluate the different types of anchoring found: exact matches, broader matches, narrower matches and evaluate the influence of these parameters
- Evaluate the query expansion's results with other types of WordNet relationships
- Use other methods for adding relationships to a thesaurus and evaluate their respective contribution

Thank you!

References

- 1. Efthimis N. Efthimiadis. Query expansion. Annual Review of Information Systems and Technology (ARIST), 31, 1996.
- 2. G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. Communications of the ACM, 30(11), 1987.
- 3. Vorhees E. Query expansion using lexical-semantic relations. In W. B. Croft and C. J. Rijsbergen, van, editors, Proceedings of the 17th Annual International ACM SIGIR Conference

Skos representation of a thesaurus

Term: Economic cooperation
Used For: Economic co-operation
Broader terms: Economic policy
Narrower terms: Economic integration, European economic cooperation, European industrial cooperation, Industrial cooperation
Related terms: Interdependence
Scope Note: Includes cooperative measures in banking, trade, industry etc., between and among countries.