Concept-Based Computer-Aided Link Generation for Electronic Handbooks

Christof Monz Jon Ragetli Maarten de Rijke
Institute for Logic, Language and Computation (ILLC)
University of Amsterdam
{christof, ragetli, mdr}@science.uva.nl
www.illc.uva.nl/~{christof, ragetli, mdr}

overview

- collaboration between Elsevier Science Publisher and the University of Amsterdam
- the aim of the project is to gain insight into setting up, maintaining and dynamically extending publication products involving multiple authors that work at a variety of levels and with a variety of backgrounds, with the overall goal of mapping out a scientific discipline
- the traditional format: handbook

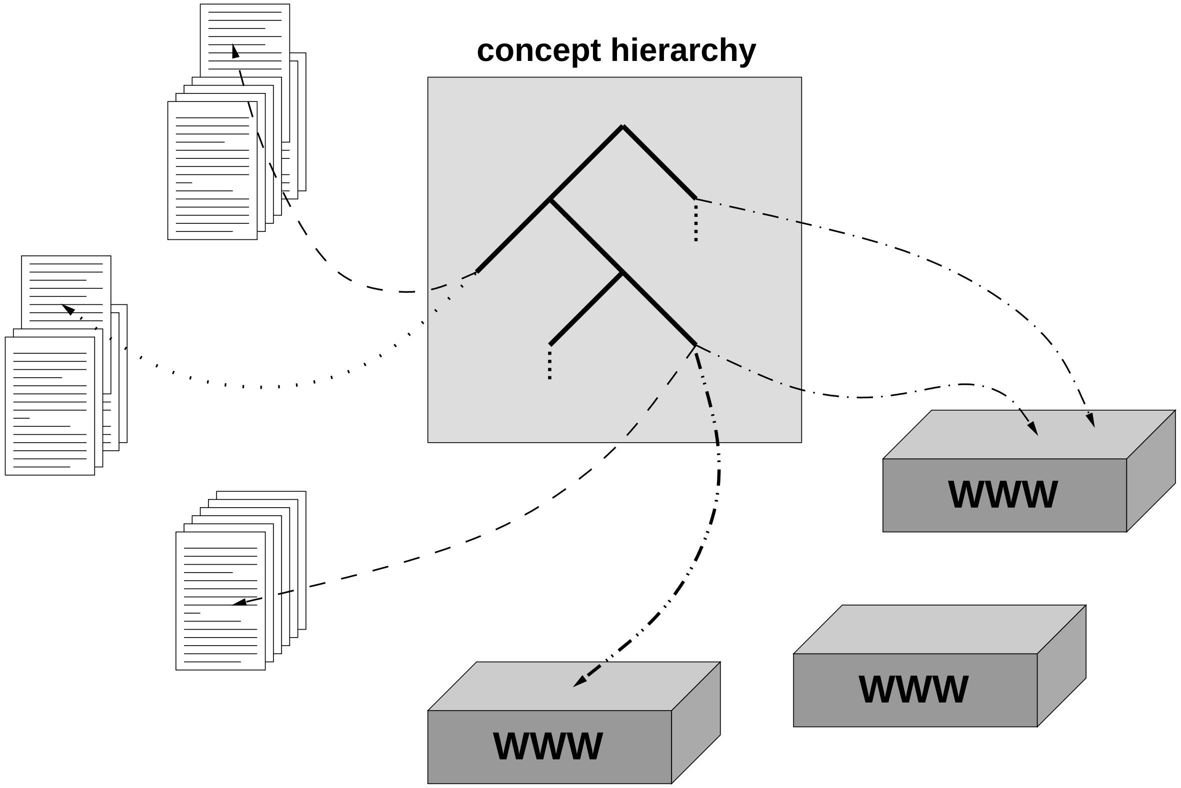
starting point

- Handbook of Logic and Language, edited by Johan van Benthem and Alice ter Meulen, 1997
- ‘interface of logic and natural language, including its broader ramifications in computer science, linguistic theory, and cognitive science’
- 20 chapters, 32 authors

current state of the project

- strategic decisions
  - use of a glossary (or concept hierarchy) to aid navigation
  - link entries to both specially commissioned 'chapters' and broader information sources
- two lines
  - developing and maintaining the glossary
  - linking the glossary to the chapters and information sources

architecture



concept hierarchy ...

- nodes in the hierarchy are terms (originally from the handbook's glossary together with their explanations)
- each head of a glossary entry is augmented by a list of synonymous terms
- a parent term refers to a more general concept than its child: the parent's concept subsumes the child's
- the child covers a related sub topic of the parent
- forming a strict hierarchy, where every child has only one parent, is not important; the subsumption structure is a directed acyclic graph (not necessarily a tree)

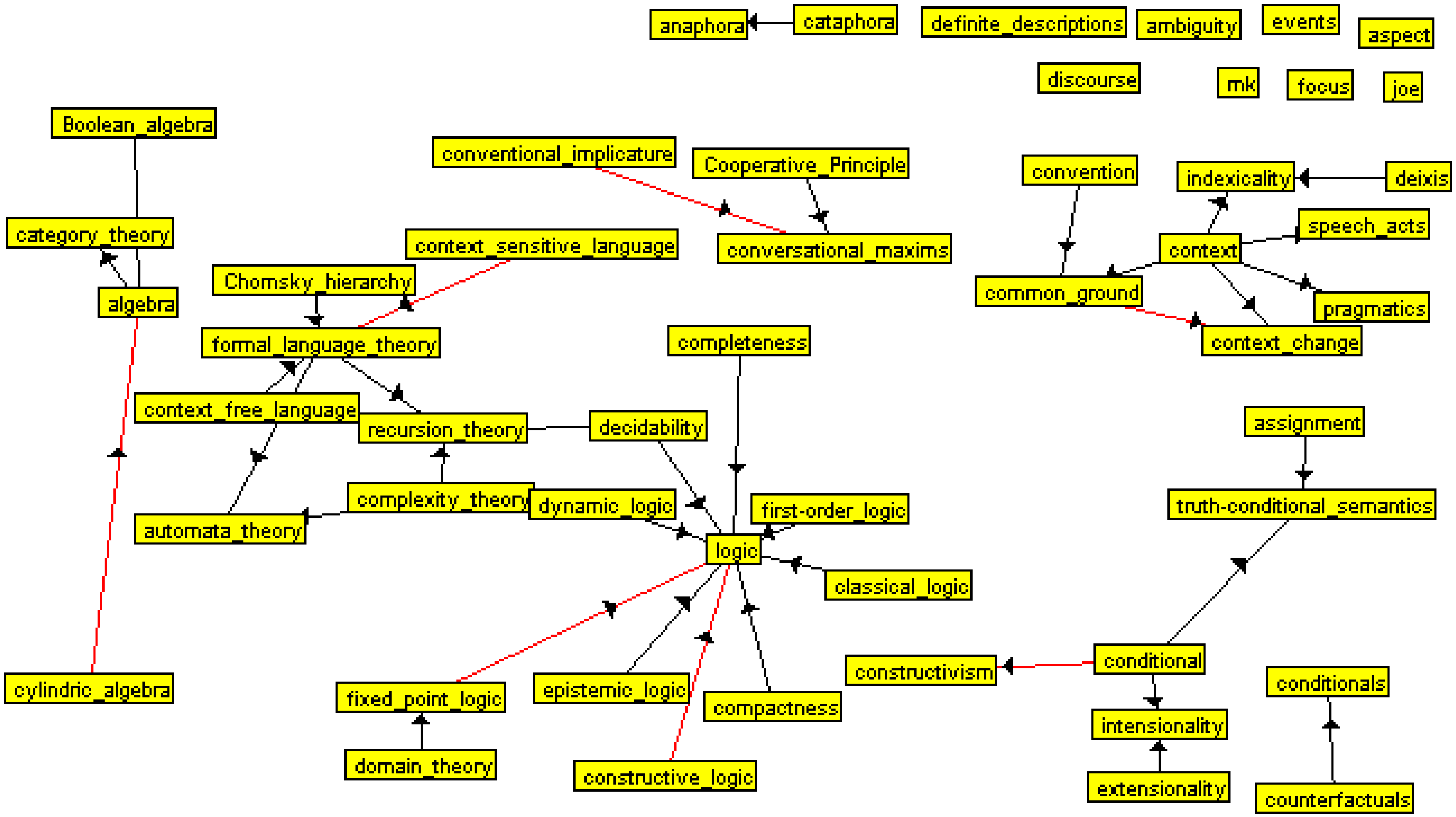
. . . concept hierarchy

- ambiguous terms have separate entries
- there is a limited number of thematic relations (such as tool, example, see also)
- currently, the concept hierarchy contains around 500 nodes

building a concept hierarchy

- the hierarchy is built by hand by a team of domain experts
- the construction is aided by automatic tools for deriving concept hierarchies:
  - coverage (recall)
  - precision
  - visualization
  - various scripts for checking structural aspects

fragment of the original concept hierarchy



types of links

currently there are four types of links:

- definition: definition of a particular glossary entry
- result: important theorems or propositions related to an entry
- related text: passages that are semantically related to an entry
- references: bibliographic entries related to the glossary entry

automated linking

- links to definitions and results employ pattern matching to identify the relevant portions of text
  - a glossary entry is linked to a definition or result if its head appears in it
- relevant text is identified by computing a similarity score between a glossary entry and a text passage
  - similarity is computed as the inner product between the glossary entry and the text passage
- related references are linked by identifying relevant text passages and looking for citations within those passages

applying IR techniques to link generation

a passage p is linked to a glossary entry g if the similarity score $sim(g,p)$ is

- above a certain threshold τ
- p is among the n top-ranked passages

experimental issues

- what is an appropriate definition of $sim(g, p)$?
- what should count as a passage?
- what should be included in the representations of g and p?
- what is an optimal value for τ?
- how many top-ranked passages should be considered?

preliminary setup

- $sim(g, p)$ is defined as the inner product of g and p where g is split into two sub-vectors g_h (glossary head) and g_e (glossary explanation):

$$sim(g, p) = \alpha \cdot (p \bullet g_h) + p \bullet g_e$$

- term weights are computed using Salton & Buckley's tfx.nfc weighting scheme
- a passage is a fixed-length unit of text (350 words)
- only words (no phrases) are indexed

evaluation of the links

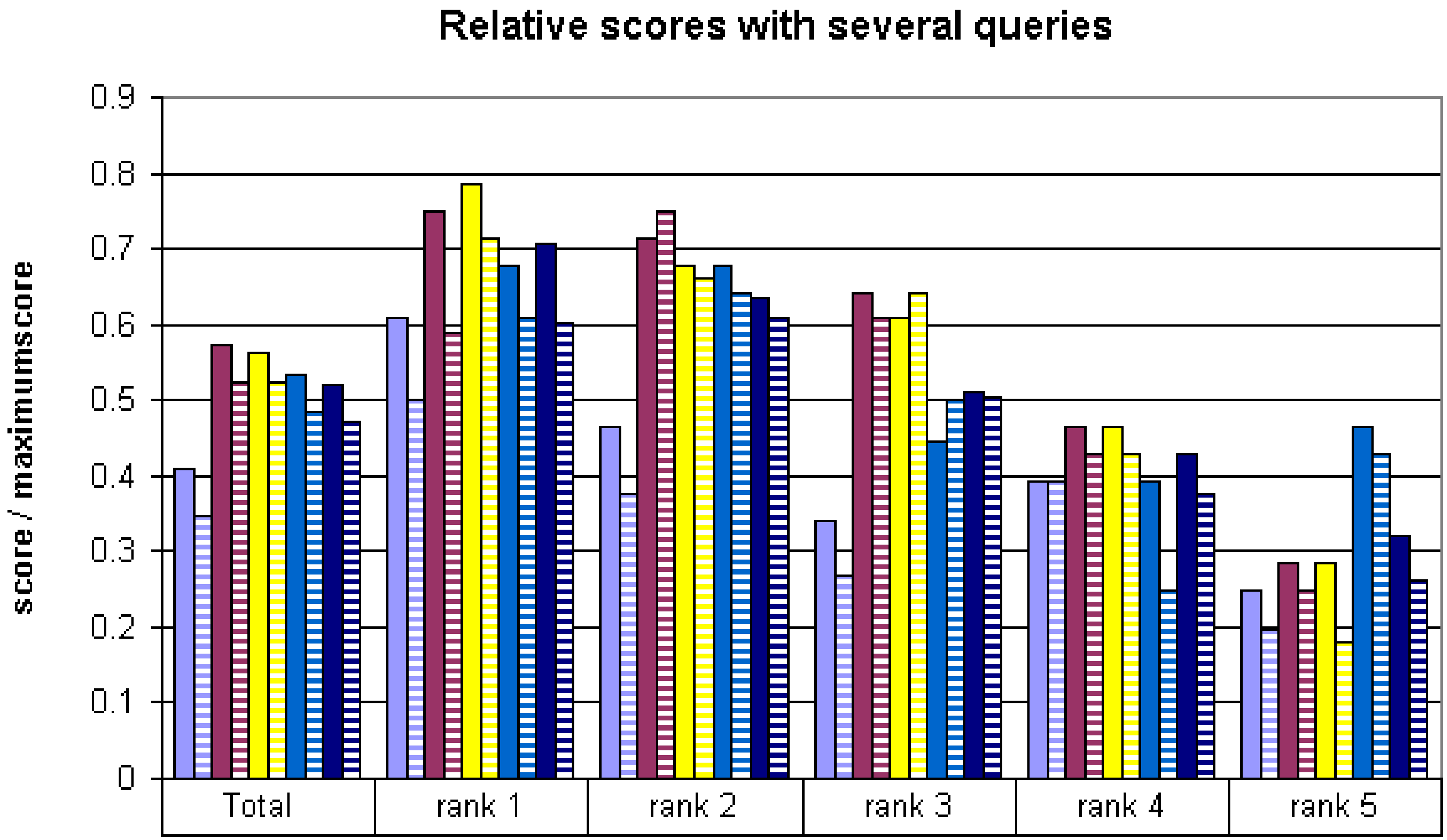
problems when evaluating automatically constructed links:

- no standard reference test collections
- inter-user disagreement on relevance of a link

in this project evaluation was done by

- arbitrarily selecting glossary entries ($\approx$ 30) and manually assessing their top n related passages (links)
- quality is defined purely in terms of precision

preliminary findings



conclusions

- the construction of an electronic handbook provides an interesting playground for testing out information retrieval techniques
- next steps in the project
  - glossary development and navigation
  - linking experiments
  - disclosing external sources via wrappers