Seeding Strategies for Semantic Disambiguation

Extended Abstract

Annika Hinze, David Bainbridge
University of Waikato
Hamilton, New Zealand
hinze,davidb@waikato.ac.nz

Rebekah Wilkins, Craig Taube-Schock
University of Waikato
Hamilton, New Zealand
wilkins.cschock@waikato.ac.nz

J. Stephen Downie
University of Illinois
Urbana-Champaign, USA
jdownie@illinois.edu

ABSTRACT

Semantic disambiguation determines the meaning of words and phrases in a text, for which we use an automatically-generated Concept-in-Context (CiC) network. Words and phrases rarely belong to a single concept; disambiguation in Capisco relies on interplay between words that are in close vicinity in the text. Starting the disambiguation is a seeding process, that identifies the first concepts, which then form the context for further disambiguation steps. This paper introduces the seeding algorithm and explores seeding strategies for identifying these initial concepts in text volumes, such as books, that are stored in a digital library.

CSC CONCEPTS

• Information systems → Information systems applications;
  Digital libraries and archives;

KEYWORDS

Semantic Disambiguation, Semantic Search, Capisco

ACM Reference Format:


1 INTRODUCTION

Semantic Disambiguation determines the meaning (semantic concepts) of words and phrases in a text to allow for subsequent semantic-enhanced search or export of semantic concepts [2]. In Capisco—a semantic disambiguation technique we have devised—the interplay between words within a certain vicinity of the text is a core feature. Further we use an automatically-generated Concept-in-Context (CiC) network that is created from analyzing Wikipedia link structures. Used together, this means the approach in Capisco can also help, for example, with alleviating issues introduced by OCR-based digitizations, such as used by the HathiTrust Digital Library (HTDL). Starting the disambiguation of words and phrases in a document is a seeding process that identifies the first concepts, which then form the context for further disambiguation steps throughout the document. This extended abstract introduces the seeding algorithm and explores strategies for identifying these initial concepts. Note that Capisco uses neither machine learning nor natural language processing for disambiguation, but rather relies on co-occurrence of semantically-related words.

2 SEEDING FOR DISAMBIGUATION

The disambiguation process has two phases: seeding and full semantic analysis. Both phases use the CiC network, in which Capisco stores triples of literal, concept and context information. The literals refer to words or phrases (e.g., “Apple”), the concepts refers to the meaning of the literal (e.g., [Fruit] or [Computer]), and the contexts refer to the circumstances under which a literal refers to a concept (e.g., [Nutrition] or [IT]), see Figure 1. The above example is described by two triples. The full semantic analysis of terms in a document (introduced in [2]), analyses the presence of other terms in the vicinity. These may, in turn, also be ambiguous but a shared context of these terms may allow for disambiguation. This shared context is determined by the initial seeding.

The seeding for the full disambiguation in Capisco starts by analyzing the stream of tokens from a corpus of text. We focus here on those produced by the OCR analysis of a text. Using the CiC network, Capisco identifies the first 100 terms, i.e., the first 100 phrases or words that can be found as literals in the CiC network. The next step determines the seed concepts by testing if any of these terms have loose semantic connections (vs the stronger connections identified and relied upon in the full semantic analysis). To return to our previous example: let’s look at the two terms “Apple” and “Autumn” selected from the 100 terms. We seek to find if they have a loose semantic connection in the CiC network. This means that the concept [Apple] for the first term “Apple” acts as context for the concept [Autumn] and vice versa. This was identified by
3 SEEDING STRATEGIES

Capisco’s seeding process as described above has successfully been used for page-based disambiguation (seeding per page). As this creates a large overhead and does not carry forward any previously found seeds, it is most appropriate for documents in which each page contains distinct information (e.g., poetry, encyclopedias). For volume-based analysis discussed here, the seeding process starts at the title page.

**Varied seed base:** Titles often contain terms that are strongly represented in Wikipedia, which leads to a narrow set of seed concepts as they all circle around very similar topics (e.g., 9 different grammars and 12 languages among 30 seeds) for [3]. Instead of allowing each seed term to participate in more than one semantic connection, we progressed through the seed terms once a match had been found, which led to a greater variation of seed concepts.

**Extracted Features:** We used the 100 most significant nouns of each volume as seed terms, based on Part-of-Speech tagging on unigrams (enforces copyright protection) [4]. Term significance was identified using AntConc [1]. Unfortunately, the restriction to unigrams prevents recognition of significant terms such as "Isle of Man" [3] or person names.

**Metadata:** We used curated volume metadata [4] as seed terms, combining title, author and publisher information. We added any names that were provided in normalized first-name surname form. As this data does not contain any OCR errors and with names matching Wikipedia’s naming convention, this strategy provided best results.

4 CONCLUSION

This paper explored seeding strategies for volume-based semantic disambiguation in digital libraries based on the Wikipedia linking structure. Our test results for a set of 100 out-of-copyright volumes from the HTDL is available on github (https://git.io/vN9iE).

REFERENCES


Figure 2: Example of two seed concepts [Apple] & [Autumn]