

Automated Document Annotation at WUR Library

Context and problem

Wageningen University and Research Centre (WUR) is as an internationally leading education and research organisation, in particular in the domain of food production, nature and climate. The task of the WUR library is to offer students and scientists the literature they need for their studies. The library acquires, stores, and provides access to a variety of specialist documents tailored to their user's needs.

The size of the document sets cause problems in terms of access. It becomes increasingly hard for users to find the right documents. One of the strategies the library employs to combat this problem is annotation with a thesaurus. Experts employed by the library determine for each new document which concepts are important to describe the resource with. The annotations help the search algorithm to produce relevant search results. The CAB thesaurus is used for this purpose, one of the few professional agricultural thesauri that exist.

Not all document sets have annotations. The Wageningen Yield database (WaY) contains publications by WUR employees which were never annotated. Annotating these documents by hand requires a considerable investment. Automated or semi-automated annotation becomes an attractive alternative. The problem is if and how automated indexing can deliver high-quality annotations like human experts do.

Approach

The approach to the problem is to develop software for automated annotation for the WUR library context. This can be based on existing annotation software or on existing software libraries. Some of these systems employ simple text matching techniques, others use machine learning techniques or features of the thesaurus. Several techniques will be tested to see which is more viable, and an analysis is made of the performance of each technique. Based on the factors that influence performance a technique can be chosen.

Some specific problems that the techniques should attempt to solve:

- some terms can mean different things than the meaning in the thesaurus; e.g. "Landscaping" and "horticulture" refers to small gardens in GB/US, while here they apply to large areas
- the structure of the CAB thesaurus may not be suitable for the techniques. The proposal should then outline how the thesaurus needs to be changed, and what benefits can be expected from these changes.

Testing of techniques will be based on a "benchmark" or "golden standard". A set of documents is first annotated by hand. These annotations are considered as complete and correct annotations. The annotations generated by the techniques is compared to the golden standard to evaluate their performance. A "baseline" technique is used to determine the lowest performance one can expect. Typically a baseline uses a simple

heuristic. The "smart" techniques need to show if and how they can improve on the baseline.

There are some issues with this type of evaluation. For example, some of the generated annotations may not match the hand-made annotations, but are not entirely wrong either.

The thesis should outline the follow-up steps that the library can make, e.g. adapting the thesaurus in a specific way to improve performance.

Research questions

- is automated annotation a viable alternative to hand-made annotations?
- how can we evaluate performance in an adequate manner?
- can customization of the software significantly boost performance?

Requirements

The student should have sufficient programming experience (in Java), as a vital component of the project will be to program. Experience with natural language processing and machine learning software is not required, but the student should be willing to become proficient in these. The student will work together with the experts at the Wageningen library. Supervision is done by Prof. Jan Top from Wageningen University and VU University, and Mark van Assem from VU University, Amsterdam. Travelling to Amsterdam for consultation will be necessary.

Contact

Please contact Mark van Assem (mark@cs.vu.nl) for more information.