



Sub-topic Annotator

OVERVIEW

The Papyrus sub-topic annotator is intended for annotation of textual news items according to subtopics of an ontology. Given an input document and a list of keyword-based patterns for each one of subtopics, the algorithm provides one subtopic name. Subtopics and their indicators can be modified independently from the annotation algorithm in a separate file which is used for automatic population of pattern matching rules.

INNOVATION

The Papyrus sub-topic annotator improves the State-of-the-Art by using an advanced annotation approach, called Cerno, based on fast and scalable compiler-style techniques: the TXL transformation framework. It only requires a small set of examples for manual specification of annotation rules.

BUSINESS IMPACT

This module can be integrated in a textual analysis component that uses the relevant ontology for recognition of subtopics. Currently it is tuned for use by the Papyrus textual analysis module using the Papyrus news ontology. In order to expand the annotator to a different application domain, the module will require a revision of the annotation rules and eventually must be tuned to another format of the input document.

INTEROPERABILITY

The core application is a TXL program that can be easily executed through a command line interface requiring a TXL interpreter for x86-based architectures (32/64bit) using either MS Windows or Linux. The module can be wrapped in a web service or called from a program written in other programming languages through command line calls.

