canarie

pacte.crim.ca

PACTE

Helping machines learn to accelerate research

Humanity is awash in a constant flood of data, and research often requires pouring through thousands of data points to look for evidence and gather statistics. While computers are excellent at some text processing tasks such as searching for keywords, they are very easily confused by human language and lack the ability to understand typos, shorthand, idiomatic phrases, archaic spellings, and most importantly, the intent behind words. Without natural language training, computers cannot sift through large bodies of text except in the simplest of cases.

This prompted researchers at Centre de recherche informatique de Montréal (CRIM) to create PACTE, a research software platform for collaborative text annotation and analysis. Training computers to understand text requires the use of annotations: small tags inserted into a text that explain to the computer what the text represents. These annotations point out how the text's grammar is structured, what meaning the text may contain, and what specific constructs are worthy of attention. PACTE is designed to simplify the entire process of machine learning: managing huge text databases, manually annotating texts, training learning algorithms, computer-annotating text, and analyzing the results.

Automating annotation and distributed collaboration

PACTE is unique in its collaborative-by-design approach. By distributing the work through a web portal, many human annotators can collaborate to divide the work on a large body of data. Once humans annotate enough of the input material, research teams can train PACTE on the resulting data. PACTE can then run algorithms to automatically annotate the remainder of the data, dramatically reducing the time for new searches and analysis.

Although machine learning is one constructive application of PACTE, the power of text annotations can have exclusively human uses too. Some texts are very short or highly specialized, leaving them unsuitable for machine learning. As well, human experts in a specific domain are often globally distributed. By allowing numerous collaborators to annotate a text, PACTE can help with highly specialized tasks such as identifying unattributed authors of seventeenth century Parisian plays or mapping out a historic explorer's travels by their journal entries.

PACTE is a research software platform for collaborative text annotation and analysis. It is designed to simplify the entire process of machine learning.

Solving real-world problems

PACTE has been used most recently to process juvenile criminal intervention reports. Social workers generate thousands of reports but in order to protect the privacy of young offenders, the reports have a limited lifespan. The volume of reports makes it difficult to perform timely human analysis before the records expire, so researchers are using PACTE to automatically detect and classify interventions with a goal of improving the quality and effectiveness of social intervention.

Reusing software for new discoveries

PACTE is part of a rich framework of software reuse supported by CANARIE. It builds upon an earlier CANARIE-supported project, VESTA (Video Annotation Processing System) that includes audio and video file transcoding and language retrieval services. PACTE also contributes a number of software services back to CANARIE's Research Software Registry, including lexical and linguistic analysis, an annotation repository, a text-file transcoding service, and a service for semantic analysis.

Platform: PACTE

Description	Helps users annotate large text corpora through a web interface. Automated services for linguistic, lexical and semantical annotations are available in both official languages. The platform also includes an active learning service that identifies the most significant data to annotate, allowing for semi-automatic annotation.
Contributor(s)	Centre de recherche informatique de Montréal (CRIM)
Research Subject	Natural language processing
Portal	patx-pacte.crim.ca
Portal Access	Free to use for non-commercial purposes upon request
Supports Separate Projects	Yes
Citizen Science	Yes
Software License	Mix of proprietary and open source
To Learn More	https://science.canarie.ca/res/115