PACTE : a collaborative platform for textual annotation

Pierre André Ménard, Caroline Barrière
Centre de recherche informatique de Montréal
pierre-andre.menard@crim.ca, caroline.barriere@crim.ca

Abstract

In this article, we provide an overview of a web-based text annotation platform, called PACTE. We highlight the various features contributing to making PACTE an ideal platform for research projects involving textual annotation of large corpora performed by geographically distributed teams.

1 Introduction

With the availability of large amount of textual data on the web, or from legacy documents, various text analysis projects emerge to study, analyze and understand the content of these texts. Projects arise from various disciplines, such as psychological studies (e.g. detecting language patterns related to particular mental states) or literary studies (e.g. studying patterns used by particular authors), or criminology studies (e.g. analyzing crime-related locations). Text analysis projects of large scale often involve multiple actors, in a distributed spatial setting, with collaborators all over the world.

While their perspectives are different and their goals are varied, most text analysis projects require some common functionalities: document selection (to gather a proper corpus for pattern analysis), text annotation (to mark actual metadata about documents, paragraphs, sentences, words or word segments) and annotation search (to search the annotated segments for the ones of interest). Furthermore, many projects would benefit from basic automatic annotation of textual components (sentences, nominal compounds, named entities, etc). Yet, each project would likely also have its particularities as to what are the important text patterns to study, and perhaps such patterns are best annotated by human experts.

We are in the process of developing a text project management and annotation platform, called PACTE (`http://pacte.crim.ca`), to support such large-scale distributed text analysis. A key component of PACTE is to not only allow for easy annotation (whether manual or automatic), but to also provide the very essential search component, to retrieve through the mass of texts, segments of information containing specific annotations (e.g. retrieving all documents mentioning a particular city). In its final state, PACTE will contain the common required project management functionalities, as well as common annotation services, but also allow for particularities (e.g. specialized schema definition).

The platform also aims at encouraging interdisciplinary collaborations, as much automatic textual analysis in the recent years is data-driven, using machine learning models which require a lot of annotated data. A known bottleneck to these supervised models is the lack of availability of annotated data. By providing a platform which makes it easy to annotate using user-defined schemas, we hope to encourage various users from various disciplines to perform annotation.

In the remaining of this demonstration note, we will show (Section 2) an example of an annotation project with definitions of the various terms used when discussing annotation projects (e.g. types, schemas, features, groups, etc). We will then highlight (section 3) the distinctive features of PACTE, mainly focusing on eight important aspects of PACTE, that it (1) is web-based, (2) handles large volumes of text for both annotation and search, (3) allows easy project management, (4) allows collaborative annotation, (5) provides some automatic annotation services, (6) allows users to define specific schemas for targeted manual annotation, (7) provides text search capabilities, (8) offers management of custom lexicon. Then, we compare PACTE to other platforms (Section 4) and we give the current state and future development of PACTE (Section 5).

2 Example of an annotation project

To give a sense of the conceptual framework underlying the development of PACTE as to provide a useful platform for large-scale text annotation projects, we choose to describe a particular use case. Let's assume a research project involving the annotation of temporal expressions (numeric or nominal expressions denoting a period or point in time) on court decision documents, which often detail the timeline of a court case. The original format of the documents (pdf, doc, rtf, txt, etc) makes up the *source corpus* while the raw text document extracted from the source would make up the *imported corpus* (also referred to as *corpus*). The researcher might also have, hopefully, research assistants available which will act as *annotators* for the project.

The researcher leading this project might choose to use the TimeML specification which includes several data structures to specify different temporal information. Let's take two of them : Timex3 for tagging explicit temporal expressions (e.g. "three months ago", "January 3rd in the morning"), and Event for expressions describing elements being positioned in time (e.g. "M. X bought his car"). Each of them, Timex3 and Event, can be viewed as an *annotation type* defined by a specific data structure, or *schema*, containing attributes and types. For example, the TimeML specification indicates that the Event annotation type includes three attributes : a unique identifier of type integer, a comment of type string (free text) and a class as an enumeration (occurrence, perception, reporting, etc.). These information would be declared as *attributes*, with their associated *data types*, in the Event schema.

PACTE would allow the definition of particular schemas such as Event or Timex3, at the start of a research project annotation task. Users would be assigned to this task in order to peruse the documents and create *annotation instances* (referred as *annotations*) for each expression they deem relevant for the Event or Timex3 type. As the researcher might want to check the agreement between annotators, *groups* of annotation instances can be created and given access to only one annotator at a time, thus preventing unwanted interactions. Schemas can also be defined at the group level, therefore allowing the same instances to be annotated by different users using different schemas. For example, one annotator might annotate using the standard Event schema in one group while another annotator in a different group would use an enriched schema with more attributes to define events. The researcher can then have access to both groups to compare annotation spans and attributes values. While all researchers will not use every aspects of these concepts (types, schemas, groups, etc.), they are an integral part of PACTE as to give a lot of flexibility to address the needs of many projects.

3 PACTE features overview

While some annotation platforms, both open-source and proprietary, already exist with a variety of features, our experience has shown some limitations when using them for large scale, multi-user annotation projects. PACTE is our humble addition to the group of available platforms in the hope to give researchers the possibility of seamlessly managing large-scale multi-user projects with minimal effort, giving them time to focus on the actual research aspects of their research. Following are some features of PACTE relevant to this goal.

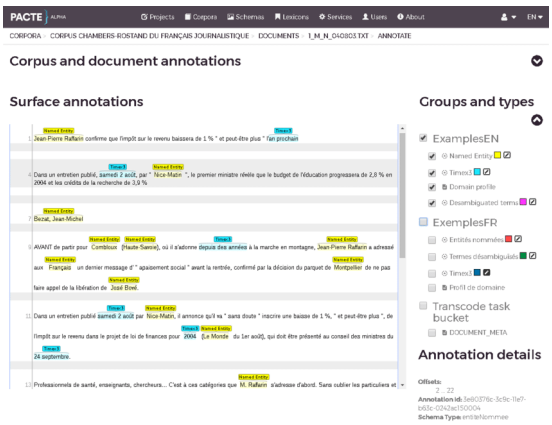

Figure 1: PACTE's main annotation window

Web-based applications such as the PACTE platform provide easy access to any user with internet access, regardless of their location, which is useful for project collaborations between research teams from different institutions. PACTE thus enables users to upload source corpora, define schemas, manage projects and annotate documents via its web interface with-

out any additional software installation or access rights
other than being able to run a web browser. The user interface is designed to easily and quickly be changed from one language to another depending on the user preference (currently in French and English, but can be expanded). In addition to the user interface, PACTE also features a secured web service (a programmatic interface) through which users, using their own credentials, can access their own raw data (documents, annotations). Following the example from Section 2, a researcher could download, using the secured web service, the Event and Timex3 annotations created by annotators on his project, train a machine-learning algorithm to link the two types of annotations (using TLINK type of annotation from TimeML standard), upload the new annotations back into the corpus using the secured web service and view the result in-situ using the user interface of PACTE.

Large corpora are often used in projects involving web-crawled documents as a source corpus or when dealing with crowd created resources like Wikipedia. PACTE architecture offers an expandable storage component in order to cope with the high volume needed by these kind of projects. Features across the platform are designed to balance the need for high volume of documents, both in the back end and front end of the platform. Furthermore, a document-level language identification service is activated at the time of corpus creation, automatically assigning a language to each document. A user can then decide to work with a multilingual corpus and keep all the documents tagged with various languages, or alternatively, the user can declare a corpus monolingual and ask the platform to discard all documents not identified with that particular language.

Easy project management is enabled in PACTE with multistep project workflow definition and management, letting the researchers plan simple or complex annotation projects with their team. For each step, the researcher can decide which corpus, annotators and schemas are used and which preexisting annotations should be seen as read-only by the annotators in order to help them with their current task (e.g. ordinals could be preannotated to help the annotation with Timex3). Projects working with large corpora will also benefit from various user-assignment strategies in order to quickly distribute multiple documents between multiple users participating in a project's step. A strategy could purposely include overlap of documents among annotators to later allow for calculating inter-annotator agreement (an important measure to be considered when annotated data is used for machine learning projects). Curation steps can also be included to validate, manipulate and prepare annotations before and after any annotation step.

Collaborative annotation is mainly done with Brat rapid annotation tool (BRAT) (Pontus Stenetorp et al. (2012)) as it contains many features desirable for PACTE. As BRAT is a complete and standalone system, the user interface[1] was unhooked from the storage backend in order to evolve with PACTE's needs. It was integrated in PACTE as part of the manual annotation window (see Figure 1) to allow the user to both easily select which annotation types to display and which annotation types to use for the current annotation task. For each annotation instance, a pop-up window is presented to the user to enter or modify values for each attribute defined in the schema.

As shown on the right section of Figure 1, PACTE features annotation groups which act as containers to manage annotations. Each single annotation must be in exactly one group and respect a schema definition dedicated to this group. These groups are used during annotation project in order to isolate the work of each annotators, simplifying management and enforcing a more rigorous annotation protocol.

Automatic annotation services are integrated into the platform in order to simplify the annotation of large corpora for researchers needing to quickly enrich text documents with various informational features. These services are provided for multiple aspects of text analysis for both French and English documents.

A first set of *linguistic* annotation services is in place to provide information on word tokens such as stem, lemma, number, genre, part-of-speech tags, or on chunked phrases (verbal, nominal, etc). These types of annotations are useful for analysis of linguistic phenomenon as well as providing character and word level information to train machine learning models.

A second set of *lexico-semantic* annotation services is designed to interact with PACTE lexicon

[1]BRAT is available at : `https://github.com/crim-ca/brat-frontend-editor`

module to tag usage of terms in a corpus. Simple string match strategies allow to tag terms' surface forms, but more complex disambiguation algorithms are included to highlight terms contextual senses as found in a multidomain lexicon (Bernier-Colborne et al., 2017). When no lexicon is available, some discovery tools can be called to extract, through statistical measures, important keywords of a document.

A third set of *semantic* annotation services targets the challenging tasks of named entity tagging, temporal expression tagging, named entity linking (wikification), to mention some of them. These services can process both English and French documents encoded in UTF-8 format.

As highlighted above, one important aspect of PACTE is to handle large corpora. As some of these services are cpu intensive, PACTE services are executed on a service gateway which manages parallel execution and load balancing. In a more advanced setup, PACTE supports process elasticity by dynamic instantiation of new processing units, enabling the system to cope with high demand, traffic or larger resources. Each service is wrapped in a container, thus easily integrating tools using different languages or technologies.

Schema definition tools will help users define schemas corresponding to their specific information needs. Each schema is split in two parts: the target and the attributes. The user first defines which entity the schema will target: a corpus, an entire document or parts of the document's textual content. Regardless of the target, a list of attributes can then be specified, each attribute being assigned a label, a data type and a description to help annotators understand the attribute's goal. Data types can be numeric, boolean, or string, either in free text or taken from an enumeration of literals, and can be declared as a single entry or an array of one of the predefined data types. This gives the user much flexibility to adapt to different types of annotation tasks. A subset of Json schemas is used internally to define each schema, providing a simple and fast format for importing, exporting and processing annotations. As this format is easy to process by both human and machine, it is an ideal balance between interoperability and operational needs. Conversion modules can be developed to adapt the contained information to other more standardized formats, like converting a text with its Timex3, Event and Tlink annotations into TimeML xml-based format.

Text search capabilities are essential when dealing with large quantities of text documents. PACTE allows for both searches on the actual text content of documents, and on their annotations (either automatically or manually generated). Although text content search is available in many text editing platforms, text annotation search is a more unique and valuable function of PACTE. For example, a user might require the list of documents containing annotations of type Timex3 with a subtype of *time*.

Lexicon management is another important module in PACTE enabling the user to define and manage a lexical resource. A lexical resource contains concepts described minimally by their concurrent terms. But each concept can be further described by a definition, usage examples, its terms genre, number, part-of-speech type, etc. Each concept can be linked to one or more user-defined domains. These domains can be defined as child or parent of other domains, thus enabling the creation of multilevel lexicons. These lexicons are made available to annotation services by snapshots in order to insure integrity of the lexical resources.

4 Platform comparison

The closest comparable project available is probably the WebAnno platform (Eckart de Castilho, R. et al. (2016)). PACTE and WebAnno share many similar features, being both web platforms, using Brat user interface as the main manual annotation tool and enabling the management of annotation projects with users, documents and annotation schemas (or layers in WebAnno).

The most discriminative aspect of PACTE is the focus on the processing of large corpora, either manually or automatically, as an annotation target. In WebAnno, corpora are managed as integral parts of projects, being stored in a SQL database on a per-project basis. Annotators select which documents they wish to annotate from the list of the project's corpus. Data is exported at the end of the project in order to be processed or aggregated offline with other annotations. Alternatively, corpora are standalone entities in PACTE as they can exist outside the lifespan of a project. Users can thus apply multiple annotation

projects (or a multistage project) on the same corpus, limiting costly manipulations over a large scale resource. Large corpora also require different selection methods of documents for the annotation tasks in order to adapt to the goal of the project, like in-depth validation or larger coverage of the resource. PACTE offers a random distribution mode to assign a specific number of documents taken from a corpus to specific users. In this mode, the project manager can also add a parameter to require that a specific number of documents should be processed by a number of annotators in order to assess their agreement. PACTE also provides a manual distribution mode to fine-tune the list of documents for each annotators, providing sort and search functions to browse through the documents of a large corpus.

Large corpora also require special consideration when applying automatic annotation tools. While WebAnno provides an automation tool that helps users annotate more efficiently, it does not offer automated annotation tools like those described for PACTE in the Section 3. Users in PACTE run these automated tools and store the resulting annotations in containers called annotation groups to keep them separate from user annotations. Access to these annotation groups can be controlled during projects to either hide or offer them as read-only resources to help the annotators. PACTE also uses a service gateway which relays processing requests on a message queue to annotation services. The service gateway manages the parallelism of multiple workers per annotation service as well as elasticity during peak periods, creating new workers to help process new requests. A scalable annotation storage service also enables the input of large quantities of automatically generated annotations.

5 Current state and road map

PACTE is currently under active development and a first partial alpha release is being tested in order to provide insights and feedback on real-world annotation projects. Some key features are being fine-tuned to improve the ease and speed at which annotators can do their work, thus improving their experience with the platform. The alpha version contains the main login page, basic corpus management (e.g. view/delete of documents), a manual annotation interface for the creation, modification and deletion of annotations. A lexicon management module is also included, as it is needed for some automatic annotation services. The next version of PACTE will include custom schema design and annotation project management, as well as automatic annotation services, as defined in section 3.

Once fully functional, PACTE will be released as an open-source project, thus enabling research groups to either host their own projects or collaborate with others on a centralized installation. Long-term plans include adding several extensions and services, such as semi-automatic annotation using active learning, corpus analysis tools, dynamic service subscription and other semantic annotation services.

Acknowledgments : This project was supported in part by Canarie grant RS-10 for Software Research Platform and the Ministère de l'Économie, de la Science et de l'Innovation (MESI) of the Government of Québec.

References

Bernier-Colborne, G., C. Barrière, and P. A. Ménard (2017). Fine-grained domain clasification of text using TERMIUM Plus. In *IWCS Workshop on Language, Ontology, Terminology and Knowledge Structures*.

Eckart de Castilho, R., Mújdricza-Maydt, É., H. Yimam, S.M., Gurevych, I., Frank, A., and Biemann, C. (2016). A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the LT4DH workshop at COLING 2016, Osaka, Japan*.

Pontus Stenetorp, S. Pyysalo, G. Topić, T. Ohta, and S. A. J. Tsujii (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Demonstrations Session at EACL 2012*.