10.12871/CLICIT2014152

ALCIDE: An online platform for the Analysis of Language and Content In a Digital Environment

Giovanni Moretti, Sara Tonelli
Fondazione Bruno Kessler
Via Sommarive 18 Trento
moretti@fbk.eu
satonelli@fbk.eu

Stefano Menini, Rachele Sprugnoli
Fondazione Bruno Kessler and
University of Trento
menini@fbk.eu
sprugnoli@fbk.eu

Abstract

English. This work presents ALCIDE (*Analysis of Language and Content In a Digital Environment*), a new platform for Historical Content Analysis. Our aim is to improve Digital Humanities studies integrating methodologies taken from human language technology and an easily understandable data structure representation. ALCIDE provides a wide collection of tools that go beyond simple metadata indexing, implementing functions of textual analysis such as named entity recognition, key-concept extraction, lemma and string-based search and geo-tagging.

Italiano. *Questo articolo presenta ALCIDE (Analysis of Language and Content In a Digital Environment), una nuova piattaforma per l'analisi di documenti storici. Il nostro obiettivo è quello di migliorare la ricerca nell'ambito dell' Informatica Umanistica integrando metodologie mutuate dalle tecnologie del linguaggio con la rappresentazione intuitiva di strutture dati complesse. ALCIDE offre una vasta gamma di strumenti per l'analisi testuale che vanno oltre la semplice indicizzazione dei metadati: ad esempio, il riconoscimento di nomi propri di entità, estrazione di concetti, ricerca basata su lemmi e stringhe, geo-tagging.*

1 Introduction

In this paper we present ALCIDE (*Analysis of Language and Content In a Digital Environment*), a new platform for Historical Content Analysis. Our aim is to improve Digital Humanities studies implementing both methodologies taken from human language technology and an easily understandable data structure representation. ALCIDE provides a wide collection of tools that go beyond text indexing, implementing functions of textual analysis such as: named entities recognition (e.g. identification of names of persons and locations within texts, key-concept extraction, textual search and geotagging). Every function and information provided by ALCIDE is time bounded and all query functions are related to this feature; the leitmotif of the portal can be summarized as: "All I want to know related to a time period".

Our work aims at providing a flexible tool combining automatic semantic analysis and manual annotation tailored to the temporal dimension of documents. The ALCIDE platform currently supports corpus analysis of English and Italian documents.

2 Related Works

Recently, several projects for the textual analysis of documents in the field of the Humanities have been presented: some of them focus only on temporal reasoning, e.g. Topotime[1](Grossner and Meeks, 2014) based on meta-data, whereas others perform word frequency analysis without a full exploitation of Natural Language Processing (NLP) techniques and temporal information, e.g. WordSeer[2] (Muralidharan and Hearst, 2013) and VOYANT (Rockwell, 2003)(Rockwell et al., 2010). Similarly to ALCIDE, WMATRIX (Rayson, 2008) is based on an automatic part-of-speech (Garside, 1987) and a semantic tagger (Rayson et al., 2004) for English to extract multiwords expressions, lemma variants and key concepts. With the only exception of key concept clouds, however, WMATRIX does not provide graphical visualizations of extracted data.

[1] http://kgeographer.com/wp/topotime/
[2] http://wordseer.berkeley.edu

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<xml archive="I.pdf">
    <file id="1">
        <head>
            <url>I.pdf_doc_number_1</url>
            <date>1901-01-01</date>
            <title>Title of the document</title>
            <description>Description of the document</description>
            <location>Geographic location</location>
            <abstract>Content of the abstract</abstract>
            <context/>
        </head>
        <content>Text of the document in italian</content>
        <original_text>Text of the document not in italian
          - if present<original_text>
    </file>
```

Figure 1: Sample of the XML input document

3 Data Preparation

ALCIDE allows the user to upload and perform analyses on any kind of documents, on condition that the documents are structured in XML according to a specific rule set.

3.1 XML Format

To fully exploit all the features of ALCIDE, the XML format must contain information about the title, the date, the location and the other information displayed in Fig. 1. A single XML file can contain multiple documents identified by a unique id. This allows users to upload an entire corpus at once.

The data can be easily imported into a database structure and given as input to NLP tools. In case the documents are available in pdf format, they need to be converted first into XML using, for instance, the JPedal PDF Java Library[3].

3.2 Data Processing

Once the documents are converted into XML and uploaded in the platform, the imported XML data is processed by TextPro[4]. TextPro is a NLP suite for Italian and English developed at Fondazione Bruno Kessler. It provides a pipeline of modules for tokenization, sentence splitting, morphological analysis, Part-of-Speech tagging, lemmatization, multi-word recognition, keyword extraction, chunking and named entity recognition(Pianta et al., 2008).

Taking the text contained in the XML file as input, TextPro returns the output of the analysis in a tabular format, with one token (and relative information) per line. When possible, TextPro modules have been tailored to the historical domain, for instance the keyword extractor and the named entity recognizer. However, we cannot expect the overall performance to be the same as for news data, on which the system was trained. The Italian POS-tagger, for instance, reached 0.98 accuracy on contemporary news stories (Pianta and Zanoli, 2007) and 0.95 on a sample of Alcide De Gasperi's writings (around 9,000 tokens written between 1906 and 1911).

3.3 Lemma Indexes

From the TextPro output, three different temporal indexes of lemmas are automatically created by ALCIDE, one for nouns, one for verbs and one for adjectives along with a timestamp. Indexing the lemma allows the portal to retrieve every document containing a certain word regardless of its declination.

3.4 Database Structure

The database structure is the core of ALCIDE and all the data presented in the graphical interface are accessible by using a query system. Data are provided both by the XML files and the TextPro analyses. The database is able to perform a large number of different queries in order to obtain the analyses requested by the user. Examples of possible queries are: the extraction of the documents published in a particular time span, in a specific city or containing a specific person name or key concept in a certain period of time. The database approach grants a good performance in case of multiple access and offers the possibility to easily update the data. Figure 2 shows that certain categories in the database such as countries or key concept can be used to group a set of documents. The database is able to relate any category to each other and then extend a category with the properties of the other related object.

4 Platform Functionalities

All processes presented in the previous Section are performed once. After the data is loaded and automatically processed, the following functionalities can be accessed through the web-based platform.

4.1 Geographical Distribution

The platform displays the geographical distribution of the documents (place of publication) and allows the user to extract all the documents related to a specific area (country or town) in a particular

[3] http://sourceforge.net/projects/jpedal/

[4] http://textpro.fbk.eu

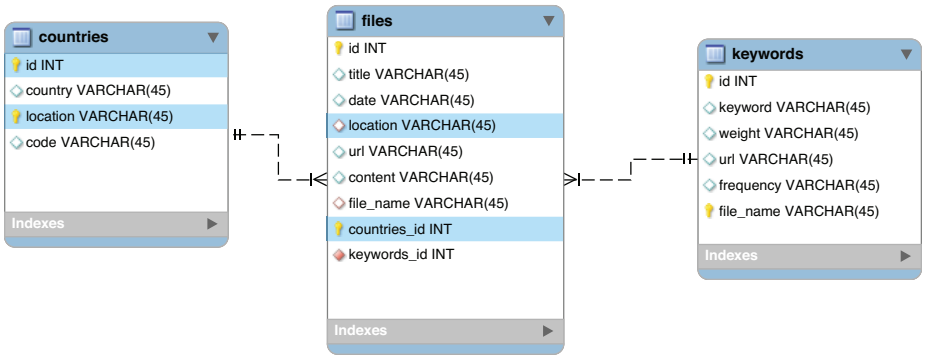

Figure 2: A simplified db graph of the structure

time span. To display data about the locations, the platform uses the Google GeoChart library[5].

4.2 Named Entity Recognition

The automatic extraction of person, location and organization names rely on the EntityPro (Pianta et al., 2008) module of TextPro. The module was originally trained on contemporary newspaper stories, on which it reached a performance of 92.12 F1 for Persons and 85.54 for GPEs. However, since the same tool obtained respectively 75.75 and 86.23 F1 on historical data (a set of Alcide De Gasperi's writings) a domain-specific adaptation was necessary. This was carried out by compiling black and white lists of common proper names for the period of interest and exploiting the tool in-built filtering functionality.

The data obtained is displayed together with the documents to highlight the most relevant persons in the text. It is also possible to query the system in order to obtain all the documents related to a specific entity or visualize in a graph the relevance of an entity over time.

4.3 Keyword Extraction

Keyword extraction is provided by the KX module embedded into the TextPro Suite. KX is a system for key-phrase extraction (both single and multi-word expressions) which exploits basic linguistic annotation combined with simple statistical measures to select a list of weighted keywords from a document (Pianta and Tonelli, 2010). KX was initially developed to work on news, patent documents and scientific articles. However, since ALCIDE is typically meant to deal with historical corpora, we tailored key-words extraction to the historians' requirement giving a higher rank to abstract concepts. This is done by boosting the relevance of concepts with a specific ending (e.g. '-ism', '-ty' in English and '-ismo', '-itidine' in Italian) usually expressing an abstract meaning. We also gave higher priority to generic key-concepts by boosting those expressed by single words.

Similarly to Named Entities, documents are displayed together with their most relevant keywords. Moreover, the portal allows the user to query the keywords characterizing a selected time span, the documents related to a specific keyword and the relevance of a keyword over the time.

4.4 Advanced Search Functions

One of the features we are interested in is to perform an efficient search of words or group of words in the whole collection of documents. The platform offers two main text search options. The first one is a full text search that gives the possibility to search for the match of one or more specific strings in a text. The second function performs a lemma based search, that looks for documents containing a specific verb, noun, or adjective in all its forms giving a lemma in input (e.g. searching for the verb *fight* the engine retrieves all the document containing *fight*, *fighting*, *fought*, etc).

Both the search functions give the possibility to perform the query in documents issued in a specific time span and to display in a graph the trend of the target term usage over time.

5 Graphical Interface

The graphical interface was developed to represent all previously mentioned data in an intuitive visualization framework. The interface provides the

[5] https://developers.google.com/chart/interactive/docs/gallery/geochart

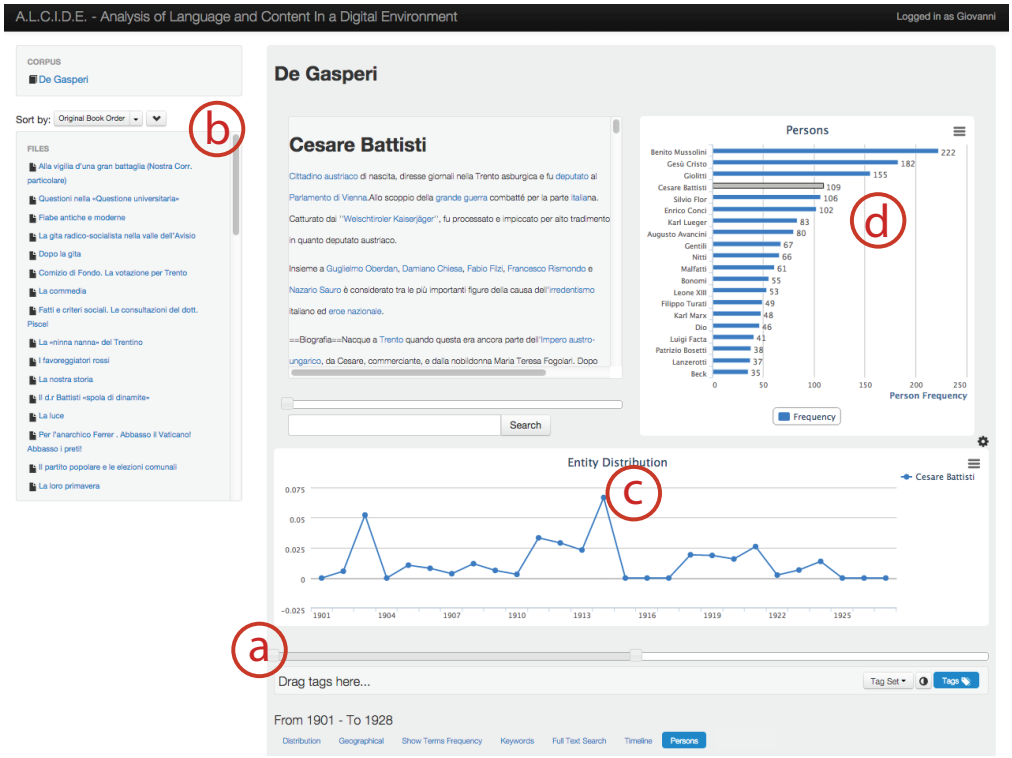

Figure 3: Sample of Graphical Interface

possibility to easily change the time span to which a search is referred. This feature is implemented through a horizontal slider (Fig.3 a) to modify the upper and the lower bound of a certain time period. The list of the retrieved documents (Fig.3 b) is always visible and accessible in the main view. In order to graphically represent the trend of an analysis (e.g. the document distribution or the number of mentions for a person) we use a line chart. All the nodes in the graph (Fig.3 c) can be used to query the system and retrieve the corresponding documents. The ranked list of keywords and entities is graphically represented by a horizontal bar chart (Fig.3 d) and are sorted by relevance to be easily identified by the user, as presented also in previous works (Few, 2013). All the bars displayed in the chart can be used to perform additional analyses by filtering and retrieving the corresponding data, for instance to get all the documents containing a particular concept in a specific time span.

Expert users can customize the set of meta-data associated with the corpus (e.g. speech transcription, propaganda materials, etc) and manually assign them to the documents. The added tags are stored in the database and can be further used to perform new queries on the collection.

6 Conclusions and Future Works

In this paper we described the general workflow and specific characteristics of the ALCIDE platform.

In the future, we aim to improve the efficiency of current functionalities and to add new ones such as (i) identification of temporal expressions and events (and the extraction of relations between them), (ii) distributional semantic analysis (i.e. quantification and categorization of semantic similarities between linguistic elements) and (iii) sentiment analysis on statements and key-concepts.

ALCIDE is already online but it is password protected. When the implementation stage will be more advanced, we will make it freely accessible and users will be allowed to upload their corpora in Excel, XML or TEI format and explore them with the platform. For the moment a video of ALCIDE demo is available at `http://dh.fbk.eu/projects/alcide-analysis-language-and-content-digital-environment`.

Acknowledgments

We would like to thank Christian Girardi for providing support in integrating and customizing TextPro.

References

Stephen Few. 2013. Data visualization for human perception. *The Encyclopedia of Human-Computer Interaction, 2nd Ed.*

Roger Garside. 1987. The claws word-tagging system. In *The computational analysis of English*. Longman, London.

Karl Grossner and Elijah Meeks. 2014. Topotime: Representing historical temporality. In *Proceedings of DH2014, Lusanne*. Alliance of Digital Humanities Organizations.

Aditi Muralidharan and Marti A Hearst. 2013. Supporting exploratory text analysis in literature study. *Literary and Linguistic Computing*, 28(2):283–295.

Emanuele Pianta and Sara Tonelli. 2010. Kx: A flexible system for keyphrase extraction. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 170–173. Association for Computational Linguistics.

Emanuele Pianta and Roberto Zanoli. 2007. Tagpro: A system for italian pos tagging based on svm. *Intelligenza Artificiale*, 4(2):8–9.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The textpro tool suite. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC)*.

Paul Rayson, Dawn Archer, Scott Piao, and AM McEnery. 2004. The ucrel semantic analysis system.

Paul Rayson. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4):519–549.

Geoffrey Rockwell, Stéfan G Sinclair, Stan Ruecker, and Peter Organisciak. 2010. Ubiquitous text analysis. *paj: The Journal of the Initiative for Digital Humanities, Media, and Culture*, 2(1).

Geoffrey Rockwell. 2003. What is text analysis, really? *Literary and linguistic computing*, 18(2):209–219.