Automatic Identification of Research Fields in Scientific Papers

Eric Kergosien, Amin Farvardin, Maguelonne Teisseire, Marie-Noelle Bessagnet, Joachim Schöpfel, Stéphane Chaudiron, Bernard Jacquemin, Annig Le Parc-Lacayrelle, Mathieu Roche, Christian Sallaberry, et al.

Automatic Identification of Research Fields in Scientific Papers

Eric Kergosien[1], Amin Farvardin[5], Maguelonne Teisseire[3], Marie-Noëlle Bessagnet[2], Joachim Schöpfel[1,6], Stéphane Chaudiron[1], Bernard Jacquemin[1], Annig Lacayrelle[2], Mathieu Roche[3,4], Christian Sallaberry[2], and Jean Philippe Tonneau[3,4]

[1]Univ. Lille, EA 4073 – GERiiCO, F-59000 Lille, France, prenom.nom@univ-lille.fr
[2]LIUPPA, Université de Pau et des Pays de l'Adour, Pau, France, prenom.nom@univ-pau.fr
[3]TETIS, Univ. Montpellier, APT, Cirad, CNRS, Irstea, Montpellier, France, prenom.nom@teledetection.fr
[4]Cirad, Montpellier, France, prenom.nom@cirad.fr
[5]LAMSADE, Université Paris-Dauphine, Paris, France, MohammadAmin.Farvardin@dauphine.eu
[6]ANRT, Lille, France, Joachim.Schopfel@univ-lille3.fr

Abstract: The TERRE-ISTEX project aims to identify scientific research dealing with specific geographical territories areas based on heterogeneous digital content available in scientific papers. The project is divided into three main work packages: (1) identification of the periods and places of empirical studies, and which reflect the publications resulting from the analyzed text samples, (2) identification of the themes which appear in these documents, and (3) development of a web-based geographical information retrieval tool (GIR). The first two actions combine Natural Language Processing patterns with text mining methods. The integration of the spatial, thematic and temporal dimensions in a GIR contributes to a better understanding of what kind of research has been carried out, of its topics and its geographical and historical coverage. Another originality of the TERRE-ISTEX project is the heterogeneous character of the corpus, including PhD theses and scientific articles from the ISTEX digital libraries and the CIRAD research center.

Keywords: text mining, natural language processing, geographical information retrieval, scientometrics, document analysis

1 Introduction

Widespread access to digital resources, via academic platforms – for example, the Gallica project (BnF)[1], the ISTEX[2] platform, electronic theses and dissertation repositories (TEL), content federation services (Isidore), or electronic publishing platforms (OpenEdition) – offers numerous possibilities for users. The ISTEX initiative was launched to create innovative information retrieval services and provide access to digital resources through different search processes. The increasing adoption of information and communication technologies by researchers in different academic disciplines, especially in the social sciences and humanities, is changing the conditions of knowledge appropriation. Digital humanities have made it possible to develop platforms, providing researchers with large volumes of academic papers and with support services to add value and make use of them (e.g., the TELMA[3] application).

The TERRE-ISTEX project was developed in this research context and proposes (1) to identify the covered territories and areas from scientific papers available in digital versions within and outside the ISTEX library, and (2) to evaluate the academic disciplines involved (e.g. history, geography, information sciences, etc.) as well as the evolution of disciplinary and multi-disciplinary research paradigms in selected topics. The results of this project will help scientists working on a given territory (areas at different scales, such as township, region, country, or continent) to retrieve papers on the same territory.

2 The TERRE-ISTEX Project

The generic approach used in the TERRE-ISTEX project is described in Figure 1. Regardless of any scientific publications corpus, a first step is to standardize textual documents. A second step is to identify, in metadata and contents of documents, the research fields as well as the scientific disciplines involved. The research field is defined as the locations constituting the territory in which the research is conducted on a given date or period of time.

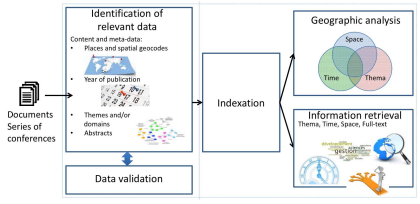

Figure 1: Generic Approach to Multidimensional Analysis of Scientific Corpus

In the experimental section, we present a Web application called SISO to help domain experts to analyze and to validate annotated text samples. The indexed and val-

[1] http://gallica.bnf.fr/
[2] http://www.istex.fr/
[3] http://www.cn-telma.fr/

idated data are then integrated into a documentary database in order to allow, on the one hand, the analysis of data and, on the other hand, the retrieval of papers on the same field and/or the same period and/or the same discipline or sub-domain using a Web-demonstrator of geographical information search.

The next section describes the corpus used within the framework of our project.

2.1 The corpus

Elaborating a corpus is a major prerequisite in the process of analysis and information retrieval. We targeted three sources of scientific papers, namely the ISTEX[4] platform, the Agritrop[5] open archive (CIRAD[6]), and a sample of PhD theses from the ANRT[7] with associated metadata available on the portal `theses.fr`.

We conducted a case study on the topic of climate change in Senegal and Madagascar. We collected an initial corpus of documents from the ISTEX platform (about 170,000 documents) using queries with the following keywords: "climate change", "changement climatique", "Senegal", "Sénégal", "Madagascar". From the same keywords, we collected 400 theses from the ANRT database. Finally, the documents from Agritrop are related to studies dealing with Madagascar and the Senegal River. The 92,000 references and 25,000 full-text documents include different types of academic papers, i.e., scientific publications (i.e., articles, etc.), grey literature (e.g. reports, etc.), and technical documentation. Each document is associated with metadata, including an abstract.

The metadata formats of the different items depend on the document origin: MODS[8] (ISTEX), XML based on the Dublin Core (CIRAD), and RDF (ANRT). The corpus is multilingual: most of the documents are either in French or in English, but there are also documents using both languages (for example, with a summary in French and a summary in English). The corpus is thus composed of multilingual and heterogeneous documents, both in terms of content and format.

2.2 An important step to standardize data

2.2.1 The process TERRE-ISTEX

Initially, the process developed by the TERRE-ISTEX project is applied to metadata and abstracts. Because of the data heterogeneity, we chose to standardize metadata using the pivotal MODS (Metadata Object Description Schema) format, recommended on the ISTEX platform. The MODS format has several advantages: (a) it is suitable for describing any type of document and any medium (digital or print); (b) it is richer than the Dublin Core; and (c) it is similar to the models for structuring bibliographic information used in libraries (e.g. MARC format). For these reasons, we apply, in a first step, an algorithm of model transformation to those 92,400 documents of the

[4] `http://www.istex.fr/category/plateforme/`
[5] `https://agritrop.cirad.fr/`
[6] `http://www.cirad.fr`
[7] `https://anrt.univ-lille3.fr/`
[8] `http://www.bnf.fr/fr/professionnels/f_mods/s.mods_presentation.html`

corpus which do not comply with this format. The second step concerns the annotation in the abstracts of spatial, temporal, and thematic entities. This step is detailed below. As a result, the MODS-TI data model expands the MODS format to describe spatial, temporal, and thematic entities extracted from documents. The MODS-TI model is detailed in the following section. Step 3 implements a new algorithm for transforming the MODS-TI format in order to create indexes so that all data can then be processed in the final stages of analysis and information retrieval.

2.2.2 The TERRE-ISTEX data model

The TERRE-ISTEX data model expands the MODS format to describe spatial, temporal, and thematic information extracted from documents and from corresponding metadata. The choice of MODS was determined by the fact that MODS is the main format on the ISTEX platform and by the advantages described above.

We added three tag sub-trees to a MODS document:

- <spatialAnnotations>,
- <temporalAnnotations>,
- <thematicAnnotations>.

In the following, we give an example of the sub-tree for spatial entities (ES). The tag <spatialAnnotations> contains a set of spatial entities (tag <es>), with the annotated text for each of them. (tag <text>) as well as its spatial footprint obtained by querying the GeoNames resource. The corresponding DTD is shown in Figure 2.

2.3 Named Entity annotation

2.3.1 Spatial entities

In the TERRE-ISTEX project, the methodology is based on linguistic patterns for the automatic identification of spatial entities (ES) Tahrat et al. (2013). An ES consists of at least one named entity and one or more spatial indicators specifying its location. An ES can be identified in two ways Sallaberry et al. (2009): as an absolute ES (ESA), it is a direct reference to a geo-locatable space (e.g. "the Plateau d'Allada"); as a relative ES (ESR), it is defined using at least one ESA and topological spatial indicators (e.g. "in Southern Benin"). These spatial indicators represent relationships, and five types of relationships are considered: orientation, distance, adjacentness, inclusion, and geometric figure that defines the union or intersection, linking at least two ES. An example of this type of ES is "near Paris".

Note that ESA and ESR integrated representation significantly reduces the ambiguities related to the identification of the right spatial footprint. Indeed, taking into account spatial indicators (e.g. "river" for "Senegal River") allows us to identify in GeoNames the right spatial footprint. To deal with distinct spatial entities with the same name (e.g. Bayonne in France and Bayonne in the United States), a disambiguation task is proposed in order to analyze the context in the textual documents Kergosien et al. (2015).

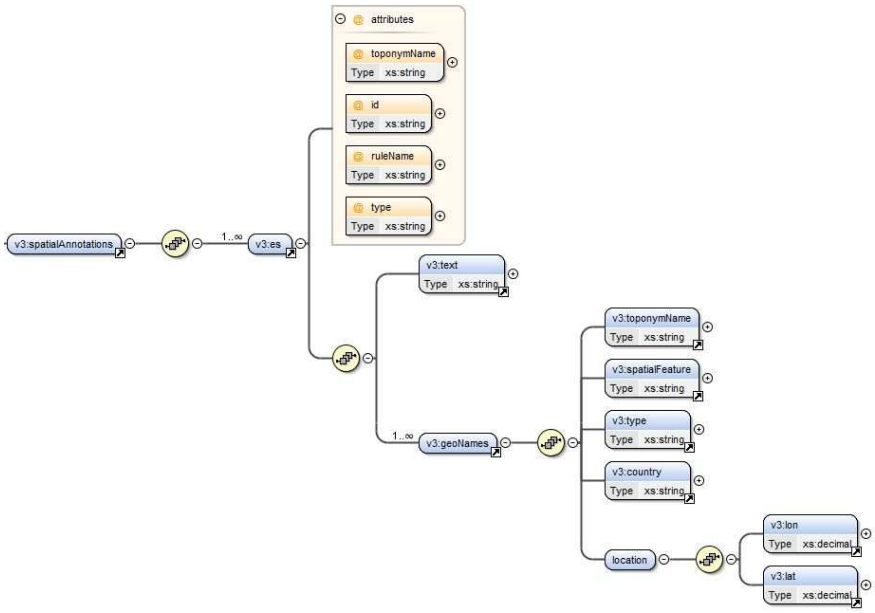

Figure 2: DTD describing the tag <spatialAnnotations>

In order to identify spatial entities, we apply and extend a Natural Language Process (NLP) adapted to Geographic Information Retrieval (GIR) domain. In this context, some rules (patterns) of Tahrat et al. (2013) have been integrated and improved to identify absolute and relative spatial entities in French (e.g. "sud-ouest de l'Arabie Saoudite" (ESR), "dans la région du Mackenzie" (ESR), "golfe de Guinée" (ESA), "lac Eyre" (ESA)). These rules have been translated in English to analyze English corpora (e.g. "Willamette River" (ESA), "Indian Ocean" (ESA), "Wujiang River Basin" (ESA)) of TERRE-ISTEX project. Moreover, we propose new types of rules in order to identify Organization (for example, an "Organization is followed by an action verb"). These different rules developed with GATE1 enable to disambiguate extracted entities and then to improve named entities recognition.

2.3.2 Thematic and temporal entities

In order to enhance the knowledge identified in metadata and to specify the subdomains, we apply modules in text mining to the content of the publications to extract domain vocabularies. First, we use domain semantic resources for lexical annotation. As, in our case, the thematic entities to be annotated are linked to climate change, we rely on the Agrovoc resource Rajbhandari and Keizer (2012). Agrovoc is formalized in XML SKOS. In the indexing phase, we mark for each term the content of an article coming from Agrovoc with the terms "used for" and the generic terms, information that will be exploited in the search engine. In the long term, we aim to propose a generic approach by giving the possibility of easily integrating a new semantic domain resource

formalized in XML SKOS. Also, we plan to integrate the BioTex module developed by the TETIS team in Montpellier Lossio-Ventura et al. (2016) combining statistical and linguistic approaches to extract terminology from free texts. The statistical information provides a weighting of the extracted applicant terms. However, the frequency of a term is not necessarily an appropriate selection criterion. In this context, BioTex proposes to measure the association between the words composing a term by using a measure called C-value while integrating different weights (TF-IDF, Okapi). The goal of C-value is to improve the extraction of multi-words terms that are particularly suitable for specialist fields.

For the temporal entities, we have integrated the HeidelTime processing chain Strötgen and Gertz (2013) to mark calendar entities (dates and periods). HeidelTime is a free, rule-based, time-sensitive labeling system for temporal expressions, available in several languages. Regarding English, several corpora of documents (i.e., scientific articles, press) have been treated Strötgen and Gertz (2013). The evaluation of this system shows better results for the extraction and standardization of temporal expressions for English, in the context of the TempEval-2 and TempEval-3 campaigns UzZaman et al. (2013) and extended to 11 languages including French Moriceau and Tannier (2013). HeidelTime produces annotations in the ISO-TimeML format, which distinguishes between four categories of temporal expressions: dates, times, durations and frequencies. Since our objective is to know the periods covered in the documents, we are only interested in temporal expressions with a calendar connotation.

3 Experiments

3.1 First experiments

The sociologists and geographers working in the project evaluated the spatial entities extraction process. The French corpus is composed of 4,328 words (71 spatial features and 117 organizations). The evaluations (with classical measures, i.e., precision, recall, and F-measure) have been investigated by comparing the manual extraction done by experts with the web service results. For spatial entities, we obtain a good recall (91%) and an acceptable precision (62%). The F-measure is 0.74. The great majority of Spatial Features (SF) are extracted but there are still some errors. The rules to identify organization are very efficient and give high precision (85%) but the value of recall is lower (67%). The F-measure for organization identification is 0.74. The rules for organization extraction seem well adapted to the domain but they have to be extended in order to improve the recall that remains low.

Moreover, in order to evaluate our approach of annotation of the ESA and ESR on a scientific corpus, we manually annotated 10 scientific articles in French, and 10 in English from the corpus on the 'climate change' topic. Items are randomly selected. The documents averaged 230 words and contained 39 spatial entities (i.e., ESA, ESR). We then annotated these documents with two processes: the CasEN chain, a reference in the field for the marking of named entities Maurel et al. (2011), and ours.

We obtained good results in terms of precision, recall, and F-measure with our process (see Tables 1 and 2). Nevertheless, we have issues with disambiguation of

named entities (Organization and spatial entities) in English and we have to improve our process for scientific articles in English.

It is significant that the results coming from the CasEN chain are far better when processing French than English too. Assuming this difference is not directly linked to the test corpus, and besides the above disambiguation issue, we consider English linguistic specificities as an explanation. In particular, links between items of a multi-word spatial entity are seldom made explicit, neither by a common morphological feature, nor by a linking preposition.

	ESA, ESR (TERRE-ISTEX)	ESA, ESR (CasEN)
Precision	100%	93%
Recall	90%	77%
F-Measure	.947	.842

Table 1: Evaluation of spatial entity annotation on 10 articles from the French corpus

	ESA, ESR (TERRE-ISTEX)	ESA, ESR (CasEN)
Precision	90%	94%
Recall	60%	53%
F-Measure	0.72	0.68

Table 2: Evaluation of spatial entity annotation on 10 articles from the English corpus

In order to validate these initial results, we are currently working on an evaluation concerning a corpus of 600 scientific articles from the ISTEX platform, 300 written in French and 300 in English. The articles were randomly selected from the 40,000 scientific articles related to the theme of 'climate change'. In this case, on the first 140 documents manually annotated by experts working in the project, we obtain a precision of 78%. The complete evaluation process is in progress. Producing a bigger annotated scientific corpus (with spatial entities, temporal entities and topics) is also an objective for other ISTEX research groups.

To help experts to analyze corpora, and particularly information related to territories, a Web application SISO[9] has been developed.

3.2 SISO Web Application to index and analyze corpora

The SISO Web application (Figure 3) allows users to upload corpora, to index documents with specific web services in order to mark different kinds of information (spatial features, organizations, temporal features, and themes), to visualize and to correct the results, and to download validated results in XML format. More specifically, it is possible to upload corpora (frame 1), each marked corpus is saved on the server and automatically available in the web application (frame 2). After having downloaded

[9] http://geriico-demo.univ-lille3.fr/siso/

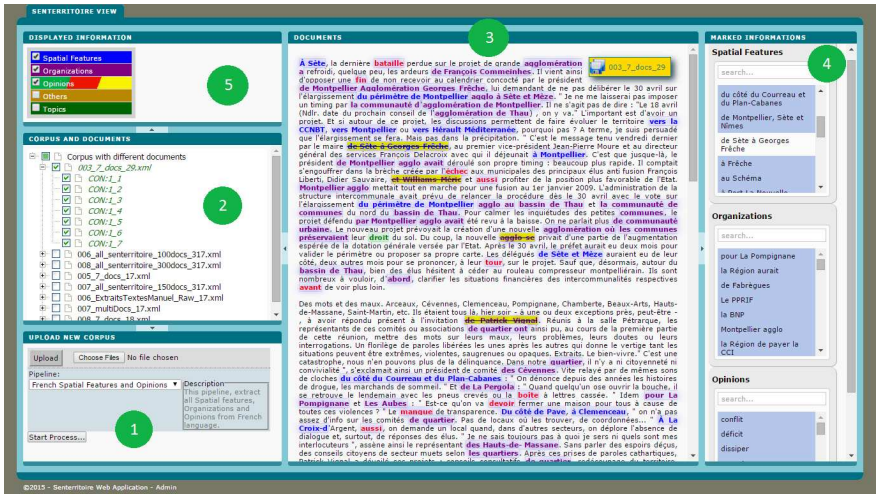

Figure 3: The Web application called SISO

documents, users can select the marked features (frame 5), see the results on the selected documents in frame 3. By selecting different categories from frame 5, the related marked information will be kindles in frame 3 and listed by type in frame 4. In case of finding any mistakes, users can unselect marked information (frame 4). Finally experts can export the selected corrected documents as a new corpus by clicking the top right button. The downloaded corpus, in XML MODS format, consists of selected documents with the marked information except those removed by the user.

The administration page allows users to upload, edit, and delete pipelines defined in the GATE format. It is also possible to remove processed corpora, to edit the uploaded pipeline rules and the available lexicons.

In order to provide experts with a web tool to process big data related to their domain, the TERRE-ISTEX approach was improved. The performance of the system, tested on 8,500 documents, are presented below:

- Temporal entity annotation: 8,196 seconds,
- Agrovoc entity annotation: 1,606 seconds,
- Search of concept and linked concept using the offline Agrovoc ontology: 36 seconds,
- Spatial entities annotation (French and English): 4,940 seconds,
- JSON index file generation: 55 seconds.

The global process takes 16,105 seconds for processing all documents, i.e., 1.9 seconds per document. This result is very encouraging.

4 Conclusion

In this article, we describe a method to deal with scientific literature on climate change from three different corpora of scientific papers. One main issue was the standardizing of data. Therefore, we have developed algorithms and a unified data model. Then, we have defined an automatic process to identify information related to a territory (spatial, temporal, and thematic information) in the documents.

Up to now, the entire corpus is indexed in JSON format. We are currently working on the enrichment of the temporal entity marking chain to integrate the BioTex tool, and on the extension of tagging assessments of marked (spatial, temporal and thematic) entities in voluminous corpus. Recently, we started to index our data with the Lucene-based search engine Elasticsearch1. Elasticsearch will facilitate the test of defined work use-cases. The main goal is to help researchers analyze big data corpora, and especially those who are interested in research related to a territory.

In our future work, we plan to use machine-learning approaches in order to improve the disambiguation process of spatial entities (i.e., ES vs. Organization). More precisely, based on our previous work Tahrat et al. (2013), we will propose to integrate the patterns described in section 3.1. as features in the supervised learning model based on the SVM algorithm. Then, we plan to compare our final model to the state-of-the-art, and specifically to the ISO-Space model produced to annotate Spatial Information from textual data Pustejovsky (2017).

5 Acknowledgements

This work is funded by ISTEX (https://www.istex.fr/), SONGES project (Occitanie and European Regional Development Funds), and DYNAMITEF project (CNES).

References

Kergosien, E., Alatrista-Salas, H., Gaio, M., Güttler, F., Roche, M., and Teisseire, M. (2015). When Textual Information Becomes Spatial Information Compatible with Satellite Images. In *Proceedings of the 7th International Conference on Knowledge Discovery and Information Retrieval*, KDIR 2015 at IC3K, pages 301–306, Lisbon, Portugal.

Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2016). Biomedical Term Extraction: Overview and a New Methodology. *Information Retrieval Journal*, 19(1-2):59–99.

Maurel, D., Friburger, N., Antoine, J.-Y., Eshkol-Taravella, I., and Nouvel, D. (2011). CasEN: a transducer cascade to recognize French Named Entities. *TAL*, 52(1):69–96.

Moriceau, V. and Tannier, X. (2013). French resources for extraction and normalization of temporal expressions with HeidelTime. In *Proceedings of the 9th International*

Conference on Language Resources and Evaluation, LREC 2014, pages 3239–3243, Reykjavík, Iceland.

Pustejovsky, J. (2017). ISO-Space: Annotating Static and Dynamic Spatial Information. In Ide, N. and Pustejovsky, J., editors, *Handbook of Linguistic Annotation*, pages 989–1024. Springer, Dordrecht. DOI: 10.1007/978-94-024-0881-2_37.

Rajbhandari, S. and Keizer, J. (2012). The AGROVOC Concept Scheme: A Walkthrough. *Journal of Integrative Agriculture*, 11(5):694–699.

Sallaberry, C., Gaio, M., Lesbegueries, J., and Loustau, P. (2009). A semantic approach for geospatial information extraction from unstructured documents. In Scharl, A. and Tochtermann, K., editors, *The Geospatial Web*, Advanced Information and Knowledge Processing, pages 93–104. Springer, London.

Strötgen, J. and Gertz, M. (2013). Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298.

Tahrat, S., Kergosien, E., Bringay, S., Roche, M., and Teisseire, M. (2013). Text2geo: From Textual Data to Geospatial Information. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, WIMS '13, pages 23:1–23:4, New York, NY, USA. ACM.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., and Pustejovsky, J. (2013). Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, USA. ACL.