

[Demo] A webtool for analyzing land-use planning documents

Mohammad Amin Farvardin^{1,2}, Eric Kergosien², Mathieu Roche^{1,3}, and Maguelonne Teisseire^{1,3}

¹ UMR TETIS (Irstea, Cirad, AgroParisTech), Montpellier, France

`amin.farvardin@teledetection.fr`

`mathieu.roche@cirad.fr`

`maguelonne.teisseire@teledetection.fr`

² GERiiCO, Univ. Lille 3, France

`eric.kergosien@univ-lille3.fr`

³ LIRMM, CNRS, Univ. Montpellier, France

Abstract. In previous work, different methods have been proposed in order to semi-automatically mine geospatial information and opinions in documents [3]. In this paper, we present the Web application, SENTIAN-NOTATOR, based on NLP methods to extract and visualize geospatial information with the associated entities. The evaluation of our application shows good results on a French corpus, i.e. F-measure of 0.74 and 0.75 respectively for the identification of spatial features and organizations.

Keywords: Land-use planning, Web application, Geospatial features

1 Introduction

Researchers and experts of land-use planning are looking for decisional tools for helping them to have an overview of user's awareness on territories. In this context, we defined the OPILAND method that enables to semi-automatically analyze sentiments related to land-use planning documents [3]. In this paper, we present the developed software and the associated web services for discovering and for integrating meaning in free texts available on the Web. This kind of textual data (e.g. blogs, newspapers, and so on) is generally complex but useful for public policy dialogue and decision-making. We thus propose an approach that enables (i) to automatically extract features related to land-use planning, and (ii) to give to experts the possibility of evaluating sentiments related to geospatial features, with the ultimate objective of evaluating the policy impact for adapting their decisions. The main originality of our software concerns the integration of different levels of semantics present in a document. This is really crucial in order to improve the analysis of information, specially for land-use planning domain. Generally in the opinion mining field, the connection between opinion and topic is studied. Actually in the land-use planning domain, it is necessary to take into account a larger number of relevant elements like spatial

features and organizations. Our software offers this possibility for helping the experts to do a finer analysis of Web data.

In this demo paper, we present the SENTIANNOTATOR web application⁴ to extract different features related to land-use planning. Hereafter, in Section 2, we focus more precisely on the deployment of natural language processing methods to extract geospatial information, i.e. spatial features and organizations. The web application for uploading, indexing, and marking textual documents is detailed in Section 3. Finally, after a quick look to our system evaluation in Section 4, future work related to our project is drawn in Section 5.

2 Geospatial information extraction

Named Entity Recognition (NER) methods identify different types of Named Entities (NE): dates, people, organizations, themes, numeric values, as well as locations. There is a significant number of available systems, such as OpenNLP⁵, OpenCalais⁶, and CasEN [6]. To recognize Named Entities several approaches are based on supervised learning methods. In this context, a bag-of-words representation is often used [1]. But this kind of statistical approach is not adapted for small data sets we are faced with in the land-use planning domain. Other approaches based on symbolic methods concern geoparsing [2, 4, 5]. The work of [5] proposes linguistic patterns to extract Spatial Features (SF) from texts. These patterns are based on a cognitive model where SF is composed of at least one NE and one variable number of spatial indicators specifying its location. Five spatial relation types are considered: orientation, distance, adjacency, inclusion, and geometric which defines union or intersection linking two SF. In our proposal, we add new patterns to improve the automatic identification of SF (absolute spatial features (A_SF) and relative spatial features (R_SF) [3, 5]). The SF annotation is based on the classical typology of the domain and more precisely on the sub-types of locations. Locations can be polysemous: human constructions (e.g. buildings) and addresses (e.g. streets). To take into account all these language specificities, some rules (patterns) have been added.

Moreover we propose a new type of patterns to identify Organizations (OE) which is a specific NE useful for land-use planning domain. The addition of specific rules enables to identify OE which could be confused with SF in documents. Such rules are: (1) an OE is followed by an *action verb*; (2) an OE is preceded by prepositions: with, by, for, on behalf of, etc.

In order to manage these geospatial information, we developed the web application SENTIANNOTATOR (<http://siso.teledetection.fr/viewer.jsp>). A screenshot is presented Figure 1. The web services use the Gate system⁷. After uploading a corpus (in French for this current version), *Spatial Features* and *Organizations* are extracted using the implemented rules. Moreover other concepts are

⁴ <http://siso.teledetection.fr/>

⁵ <https://opennlp.apache.org/>

⁶ <http://www.opencalais.com/>

⁷ <https://gate.ac.uk/>

extracted using Gate: (i) *thematics* based on a lexicon using Agrovoc thesaurus⁸, (ii) *Opinions* related to land-use planning domain [3].

3 The SentiAnnotator Web application

The web application (Figure 1) allows users to upload corpora, to index documents with specific web services in order to mark different kinds of information (spatial features, organizations, opinions, and themes), to visualize, 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 documents, users can select the marked features (frame 5), see the results on the selected documents in frame 3. In this frame, spatial features are in blue color, organizations in purple color, the positive opinions in green color, negative ones in red color and neutral in yellow color. 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 expert can export the selected corrected documents by clicking the top right bottom. The downloaded corpus consists of selected documents with the marked information except those were 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 and to edit the uploaded pipeline rules and the available lexicons.

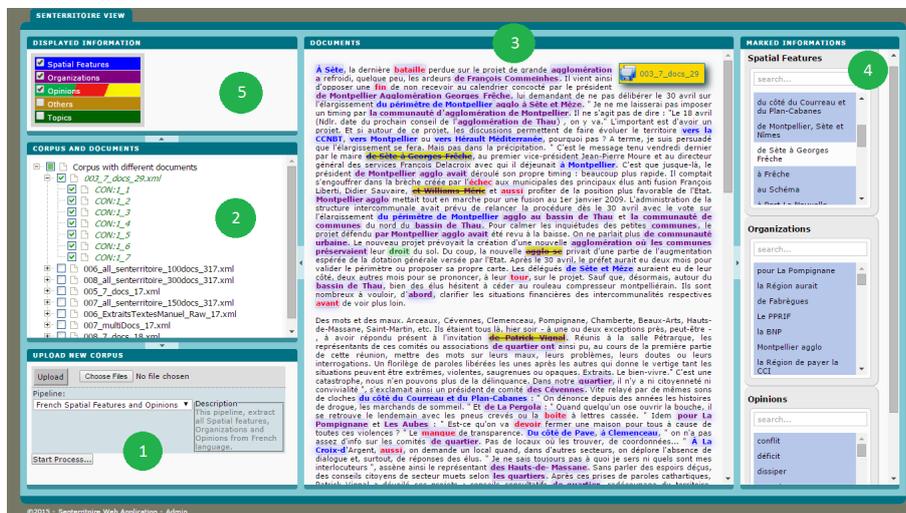


Fig. 1. SENTIANNOTATOR Web application

⁸ <http://aims.fao.org/fr/agrovoc>

4 Evaluation

Three experts of the project evaluated the process for extracting geospatial information by using SENTIANNOTATOR application. We use a French corpus composed of 4328 words (71 spatial features and 117 organizations). The evaluations (with classical measure, i.e. Precision, Recall, and F-measure) have been investigated by comparing the manual extraction done by experts with the web service results. For SF, we obtain an excellent recall (0.91) and an acceptable precision (0.62), the F-measure is 0.74. We extract the great majority of SF but the rules still return some errors. The rules to identify OE are very efficient and return high precision (0.85) but the value of recall is lower (0.67). 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.

5 Conclusion and Future Work

In this paper, we have presented a Web application called SENTIANNOTATOR including web services (1) to annotate corpora with features related to land-use planning, and (2) to evaluate achieved approaches with experts. Experts are using this tool for analyzing the construction project of a road around Villeveyrac (France). Future work will be dedicated to the improvements of the defined linguistics patterns for discovering NE in order to tackle the issues related to the land-use planning specificities and the multilingual aspects. We also plan to extend our approach to different types of textual contents such as tweets.

Acknowledgments: The authors thank Midi Libre (French newspaper) for its expertise on the corpus and all partners of the SENTERRITOIRE project for their involvement (MSH-M, GEOSUD Equipex, NUMEV Labex, and TECTONIQ PEPS project).

References

1. X. Carreras, L. Marquez, and L. S. Padró. A simple named entity extractor using adaboost. In *In Proceedings of CoNLL-2003*, pages 152–155, 2003.
2. M. Gaio and V. Nguyen. Towards heterogeneous resources-based ambiguity reduction of sub-typed geographic named entities. In *Int. Conf. of GeoSpatial Semantics*, pages 217–234, 2011.
3. E. Kergosien, B. Laval, M. Roche, and M. Teisseire. Are opinions expressed in land-use planning documents? *International Journal of Geographical Information Science*, 28(4):739–762, 2014.
4. J. L. Leidner and M. D. Lieberman. Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, 3(2):5–11, July 2011.
5. J. Lesbegueries, C. Sallaberry, and M. Gaio. Associating spatial patterns to text-units for summarizing geographic information. In *Proceedings of ACM SIGIR 2006. Geographic Information Retrieval, Workshop*, pages 40–43, 2006.
6. D. Maurel, N. Friburger, J.-Y. Antoine, I. Eshkol-Taravella, and D. Nouvel. Casen: a transducer cascade to recognize french named entities. *TAL*, 52(1):69–96, 2011.