

Semantic Annotations for Digital Investigations

Nikolaos Lagos

Xerox Research Centre Europe
6, chemin de Maupertuis
Meylan, France
+33 (0)4 76 61 51 92

Stefania Castellani

Xerox Research Centre Europe
6, chemin de Maupertuis
Meylan, France
+33 (0)4 76 61 50 72

Aaron Kaplan

Xerox Research Centre Europe
6, chemin de Maupertuis
Meylan, France
+33 (0)4 76 61 50 35

nikolaos.lagos@xrce.xerox.com stefania.castellani@xrce.xerox.com aaron.kaplan@xrce.xerox.com
com

ABSTRACT

In this paper, we describe how we believe semantic annotations can help in the provision of support for digital investigation activities for analysts searching for information in large document collections. In particular, we are interested in developing tools to support the activities of lawyers in corporate litigation. In current applications, information such as key characters and events and relations among them has to be mostly manually identified and collected from the collection of documents. We describe here our current and future work on the construction of a system that provides tools and mechanisms for helping users better extract, navigate, select and store such information. We also discuss interesting challenges that we have identified during this work.

Categories and Subject Descriptors

H.5.2 User Interfaces: User-centered design, I.2.4 Knowledge Representation Formalisms and Methods

General Terms

Design, Human Factors, Languages.

Keywords

Information extraction, natural language processing, digital investigation, legal case building and reasoning, knowledge representation.

1. INTRODUCTION

We are interested in developing tools to support the information seeking and sensemaking activities of (groups of) analysts working on large document collections. Examples of this kind of activities are lawyers working on the construction of a legal case for a corporate litigation, journalists reconstructing a series of events from documentary evidence, or a team of analysts investigating a conflict of interest or conducting an audit in an organization. In all these activities it is quite typical to have analysts searching large volumes of data looking for information supporting their line of inquires. In particular, in litigation, lawyers' work involves the analysis of large volumes of documents to find evidence for the litigation claims. Litigation involves a number of stages with different support requirements,

including the collection of all documents possibly relevant for a case, the selection of documents meeting specific criteria of relevance, and case construction, where arguments around facts and evidence are put together for presentation in court. The primary goal of the searching and browsing facilities offered in current litigation tools is to find relevant *documents*, often using keyword/boolean based search techniques. However during case construction the emphasis shifts from finding documents to finding *entities* and *actionable information* ([2], [4]) derived from these entities. Identifying the important information is time consuming and costly and in recent years there has been a move to bring into play language processing technology. However, this mainly concerns early stages of the process and limited support is offered for case building and reasoning. In particular, in current applications, information such as key characters and events and relations among them is identified and aggregated mostly by hand.

We are constructing a system that aims at helping users better extract, navigate, select and store such information. Our approach for the extraction of information is based on natural language processing (NLP) techniques and it enables the use of information about the relations among the key players of a case, extracted in the form of events [1]. (We use the term "event" broadly, to encompass not only point-in-time events, but also states, processes, etc.) In order to describe and structure the data extracted we have defined a knowledge model (or ontology), which contains concepts that we believe will be broadly useful in digital investigations. For instance, people and organizations are typical examples of *characters* that may have a role in a legal case. As we are particularly interested in the relations between these entities, special attention is given to the representation and analysis of events and how their definition influences the representation of the rest of the entities extracted. For instance, the role of the characters in a case is determined, among other factors, by the events in which they participate. Naturally, this is a two way relation. The events that a key character participates in may be important for the case and the participants of a key event may be key characters. One of the core requirements is therefore identifying other factors (in addition to the participants) that make an event important. These include: the type of an event, the role of a character in the event, the relative time of an event in the chronology of the case, and the location where the event took place.

We also believe that such structured representations can help us develop mechanisms that will allow users to better navigate and utilize extracted and inferred information. We are therefore also working on an interaction mechanism that allows the users to formulate questions in a controlled way. The mechanism is tuned for the kinds of questions that typically arise in digital

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ESAIR workshop '10, October 26–30, 2010, Toronto, Canada
Copyright 2010 ACM 1-58113-000-0/00/0010...\$10.00.

investigation activities, e.g. those conducted by lawyers or journalists, focusing on characters involved in events that may take place at a particular time and place and may be related to other events or characters.

In the rest of the paper we describe how we have used semantic annotations in the system we are building, other aspects where we believe semantic annotations could be useful, and issues that will need to be resolved before the system is ready to be put to use.

2. CHALLENGES

Digital investigation has all the characteristics of an exploratory process. There is a continuous interaction between human and machine. The machine helps to filter and detect possibly important information, and the human validates and constructs new information based on the extracted data. This is an iterative process until no more facts can be derived. Several challenges arise when designing and implementing techniques that combine extraction of information and inference of new facts coupled with interaction mechanisms that allow digital investigator(s) to create and organize their own information space. We discuss some of them here.

From semantic annotations to knowledge representation

Merely annotating individual mentions of characters and events may enable a certain amount of new functionality, but there is more to be gained by recognizing that the same characters and events are mentioned multiple times in a single document and across multiple documents, and synthesizing richer representations that combine information from multiple sources. To this end, we express our semantic annotations in a knowledge representation language, and store them in a knowledge base. We can then use inference mechanisms to identify when multiple annotations refer to the same thing. We have already implemented simple coreference resolution mechanisms for mentions of persons, but this is only a start. The mechanisms could be enhanced to integrate encyclopedic knowledge from external sources (e.g. knowing that a referring expression “he” can’t be coreferent with a name if the person with that name is known to be female), and need to be extended to other types of entities such as organizations and places. Furthermore, coreference mechanisms for events are needed. Promising directions towards that objective were set by recent international competitions and tasks [3], but many technical problems remain to be solved. There are also interesting questions of interaction design: can the user help in this process, thus adopting a semi-automatic approach instead of an automatic one, and how should we define the boundary between the system’s initiative and that of the user

Inference

Inferring that different textual mentions refer to the same real-world entity is but a special case of the more general challenge of inference. Inference mechanisms can allow the system to discover facts that are implicit in the documents. An inference mechanism could offer more automation in the following activities: help users detect new elements for a line of inquiry or possible directions to be followed; show links that can constitute evidence for a line of inquiry; detect contradictions; and identify incomplete information. Again, there are issues of both technical infrastructure and interaction design.

Integrating user-provided information with information extracted from documents

In the current version of the system that we are building we use only automatically-generated annotations, but we believe that it could be useful for the users to be able to add relevant information that is not present in the documents, or is present but difficult for the system to extract. A number of challenges are associated with such a process: should the users be able to express themselves in natural language or based on a controlled approach? User-provided annotations could be useful knowledge to be reused by the system for inferring new facts. However, in this case, where does the responsibility of the system stop and the responsibility of the human begin in terms of provenance and trust? How do we keep manage conflicting information that could be added by different users? Do we support multiple users and collaborative workflows and how?

Domain adaptation

Digital investigation can incorporate different application domains at different times (for example in the litigation case that we investigated, medical and chemical related information models should be available to help structure the knowledge in an optimal way). Where do we find the necessary world/external knowledge? Can initiatives such as Linked Data help in this direction?

3. CONCLUSIONS

We believe that semantic annotations, both generated automatically and entered manually by humans, could be useful as the basis of software support for digital investigation. We are exploring this idea using an approach that simultaneously considers technical issues and issues of interaction design, keeping in mind the strengths and limitations of infrastructure components (NLP tools, inference engines) and of human-computer interaction. The question of how to evaluate such a system is an important one: IR metrics such as precision and recall are not applicable to a system that interactively helps a user build complex representations. We will need to define new metrics and perform studies with users in the loop.

4. REFERENCES

- [1] Lagos, N., Segond, F., Castellani, S., and O’Neill, J. 2010. Event extraction for legal case building and reasoning. To appear in *Proc of IIP’10*. (Manchester, UK, Oct., 2010).
- [2] Noel, L. and Azemard, G. From semantic web data to inform-action: a means to an end. In *Proc. of CHI’08*, Florence, Italy, ACM, Apr. 2008.
- [3] Ruppenhofer, J., Sporleder, C., Morante, R., Baker, C., Palmer, M. *SemEval-2010 Task 10: Linking Events and Their Participants in Discourse*. NAACL-HLT Workshop on Semantic Evaluations, Boulder, Colorado, USA (2009).
- [4] Sheth, A., Arpinar, B., and Kashyap, V., 2002. Relationships at the Heart of Semantic Web: Modeling, Discovering, and Exploiting Complex Semantic Relationships. Tech. Rep., LSDIS Lab, Univ. of Georgia, Athens GA 30622.
- [5] Urbani, J., Kotoulas, S., Maassen, J., Van Harmelen, F., and Bal, H. OWL reasoning with WebPIE: calculating the closure of 100 billion triples. In *Proc. of the Seventh European Semantic Web Conference* (2010).