Extraction of information on activities of persons suspected of illegal activities from web open sources

Christian Fluhr, Aurélie Rossi, Louise Boucheseche, Fadhela Kerdjoudj
GEOLSemantics
32, rue Brancion, 75015 Paris, France
E-mail: christian.fluhr@geolsemantics.com, aurelie.rossi@geolsemantics.com,
louise.boucheseche@geolsemantics.com, ker.fadhela@gmail.com

Abstract

This work is part of the French funded SAIMSI project (Suivi Adaptatif Interlingue et Multisource des Informations). The aim of the project is to follow activities of persons suspected of illegal actions like terrorism, drug traffic or money laundering. The paper focuses on the information extraction in particular. This extraction is done in French, English, Arabic and Chinese. The information extraction is based on a deep morphosyntactic analysis. Recognition of single words, idiomatic expressions, compounds is performed and named entities are identified and categorized. Dependency relations are built, passive/active forms, negation anaphora, verb tenses are processed. Information extraction is application-independent and uses extraction rules. At this level some named entity categories can be reconsidered. This extraction is based on a large security ontology. The paper details the problems of the consolidation of the extracted knowledge at the document level. The future evaluation on WEPS-3 data is presented.

Keywords: Information extraction, illegal activities, RDF, web open source information

1. Context

This work is part of the SAIMSI [2] project funded by the French Research Agency. The aim of the project is to follow activities of persons suspected of illegal activities (terrorism, drug, money laundering, etc). If possible, the activities must be located in time and place (geochronolocalization).

The processed information is taken from open sources on the Internet (news sites, social networks, specialized sites) in 4 languages: French, English, Arabic and Chinese (Mandarin).

The extracted information is structured and coded into RDF in English whatever the source language, according to security ontology. This means that information coming from various languages can be merged into a knowledge base.

As the project concerns international activities, a special attention has been paid to the different spelling of person names originally coded in different character sets. This gives a better recall but increases the problem of homonymy resolution.

With the result of knowledge extraction, two databases are built.

A knowledge base that stores the triples allows reasoning to infer new relations and controls the consistency of the knowledge. The results of queries are presented in biographic sheets, geographic maps, timelines and graphs of relations between persons and/or organizations.

A cross-language text database is also built. It is used to control the origin of the extracted knowledge and gives the possibility for the user to interrogate on themes that have not been structured into the knowledge base.

This paper focuses on the extraction of knowledge based on the result of a deep general-purpose morphosyntactic analysis and an application-oriented knowledge extraction using rules. The incomplete information obtained from sentences is discussed. The complementation of the extracted knowledge by a local reasoning on the full document is presented.

This paper does not describe the introduction of new document knowledge into the knowledge base that contains the instances and properties already introduced from previous documents and especially the processing of homonyms. It is an ongoing work.

2. Related works

Developments made by the Joint Research Center (JRC) of the European Commission in Ispra [5] are the closest work to ours. This work (EMM) consists in gathering of news in various languages, processing of name variants in different character sets, construction of graphs of personal links and event extraction. The common work with FRONTEX [1] and the University of Helsinki is security-oriented. The aim of this project is to extract, in 7 languages, border security oriented events. Two approaches have been tested.

The first one (NEXUS) by JRC begins by a clustering that gathers texts arriving in a 10 minute window into groups relating to the same event. After a morphological analysis extraction rules are performed on the beginning of each text. They consider that the main information is in the beginning of texts and because of redundancy, lost information in a document can be extracted in the beginning of another. The other

approach is provided by the University of Helsinki (PULS) [6] and performs a morphosyntactic parsing of the full document including resolution of anaphora. This approach is close to ours but the problem discussed in this paper about consolidation of knowledge at the document level seems to be not taken into account.

A large amount of police and intelligence services use the tools from I2 (I2 base, Analyst's Notebook) [3]. Extraction and categorization of named entities done by TEMIS have been introduced into TextChart

AutoMark helping users to fill the Base.

3. The Ontology

The knowledge to be extracted is described in a security ontology .

For each person, a list of attributes has to be extracted: given names, surnames, middle names, nicknames, birth dates, birth places, diplomas, personal addresses, fixed telephone numbers, mobile phone numbers, email addresses, Web site, internet domain, etc. Actions to be extracted are: travelling from a point to another, contacts between persons and/or organizations, payment or purchase, construction of objects, emission of a message (discourse, book, mail, interview, etc.), events of the life (birth, marriage, divorce, death, funeral), family relations (brother, sister, son, daughter, grandson, etc.), interaction with police and justice (control by police, arrest, conviction by a court, release from prison, etc.), link and role with an organization, teaching organization attended. This ontology has been built after discussions with users. The main applications that have been considered are terrorism and malware production and use (especially phishing).

4. Deep morphosyntactic analysis

In order to minimize the effort to write extraction rules, we have chosen to build them on the result of a deep morphosyntactic analysis which is domain independent. That means that our extraction rules are built by linguists. There are no extraction rules built on the surface level that can be learned from a tagged corpus. This minimizes the number of rules, because they are applied to a deep representation that is independent of the surface level.

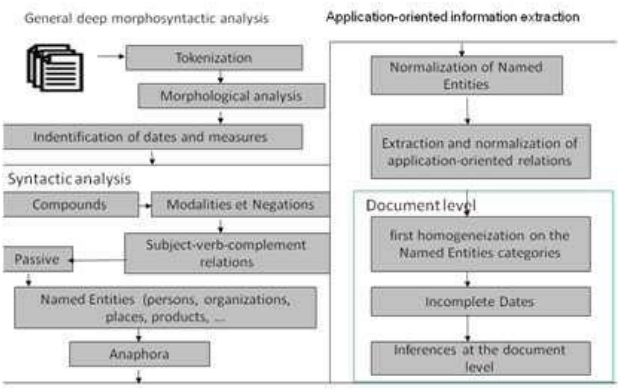

Figure 1: schema of the linguistic processing and information extraction

The morphosyntactic analysis is developed using a linguistic-oriented programming language that is executed on a weighted finite state automata engine. The morphological level recognizes words even for Mandarin language which has no space character between words. Syntactic and some semantic information is associated using a dictionary of single word forms and a dictionary of idiomatic expressions. Some named entities are recognized and normalized at this level such as dates, phone numbers, e-mail addresses, etc.

A part of speech tagging then disambiguates grammatical categories and helps to choose the lemma and more generally the normalization of each word. Normalization is a common representation of different lemmas that are full synonyms or spelling variants. It is the case for UK and US spellings like “colour” and “color”.

The syntactic parsing recognizes dependency relations inside noun and verb phrases as well as subject-verbobject relations, and gives a common representation of active and passive forms by producing agent-actionobject relations. Verb tenses including compound tenses are recognized, negation is also recognized along with modalities. Pronouns and possessive adjectives are processed. This last point is very important because without pronoun recognition a lot of knowledge cannot be extracted.

Pronouns are mainly processed at the paragraph level, but for particular type of documents like biographies it is necessary to process pronouns at the document level.

During this step, the rest of named entities are recognized and typed using both recognition rules and lists of known entities coming from databases like Geonames or DBpedia.

5. Extraction

The extraction rules are written in the same language as the one used for syntactic analysis. They are triggered using words representing different actions or properties described in the ontology. The rule verifies that it is really the supposed action and then recognizes and links different role players (agent, object, place, date, instrument, manner, etc).

Verification is necessary because some triggers can be ambiguous. For example, in French “se rendre” can be “to go” or “to surrender” that represent different actions in the ontology. Of course, if the agent is in both cases a person, the other role players have different types.

Other kinds of semantic ambiguities are processed using these rules, like “Father” which can be an ecclesiastical title or a family relation.

Example of extraction:

The result is RDF-coded but in order to minimize the place in this paper we will give the natural language reconstruction of the RDF. This natural language reconstruction of an English oriented RDF can be done in any language even if no morphosyntactic parsing is available for this language. It’s a way to have a cross-language summary of the activities interesting the users.

Original sentence *: “Basam Ayachi a organisé le mariage de Malika el Aroud avec Abd el Satar*

Dahmane qui a tué le Commandant Massoud le 9 septembre 2001.“

***Result** :*

Union organised by Basam Ayachi between Malika el Aroud and Abd el Satar Dahmane.

ViolentAct author Abd el Satar Dahmane, victim Commandant Massoud date 20010909 Death Commandant Massoud date 20010909 type: murder

The last information is not really in the sentence but is inferred from the type of violent act.

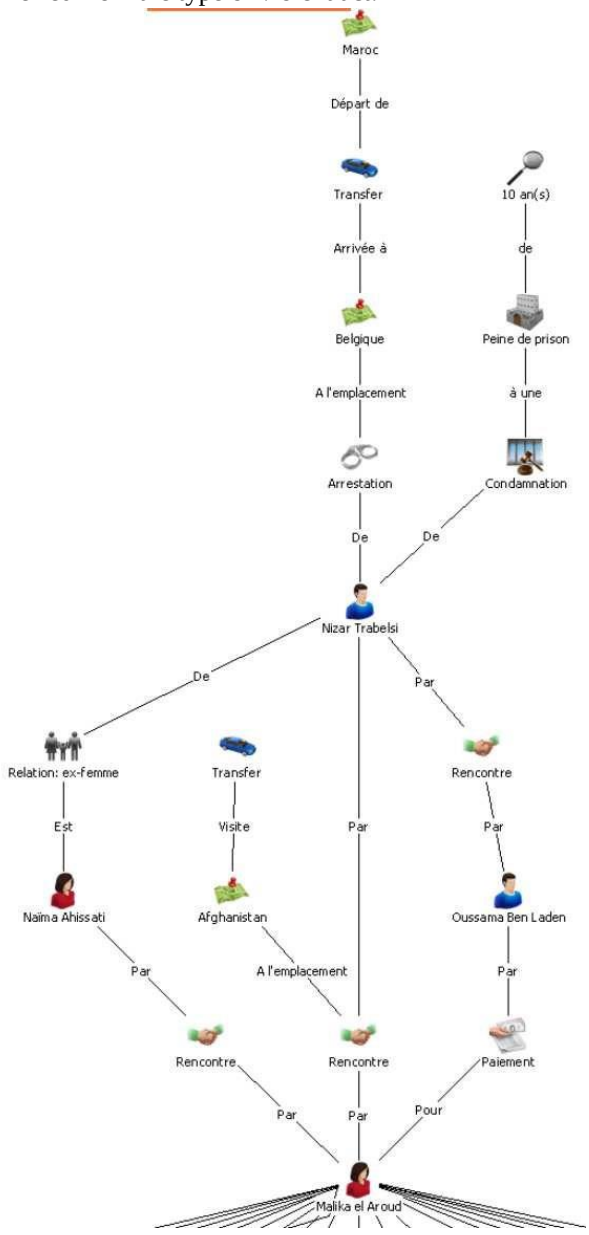

Figure 2: Visualisation of extracted information using Analyst Notebook of I2.

6. Limitation of the knowledge extraction from sentences

As you can see in the previous example, information which should be important is lacking in the sentence like the date and place of the marriage and the place of the Massoud assassination. It lacks also the given name (Ahmed Shah) of Massoud. All this information can often be obtained from the same document. In fact, the extraction must be consolidated at the document level. Documents have a semantic coherence. The discourse is constructed to prevent the reader from having confusions, but at the same time some information is not repeated because the reader can infer it.

That is the reason why it appears compulsory to use the order of succession of sentences in the documents and also the information about the publication date of the documents given by the metadata, because they are used by the reader to understand the text. **6.1 Identification of persons and places** Persons and places are cited in different sentences with sometimes different forms. If the author wants to give information about two different people having the same surname, he will prevent confusion by giving distinctive information (given name, Jr/Sr, title). On the contrary, compatible names can be gathered using the same identification number.

6.2 Relative dates

Some relative dates depend on the publication date. In this case the tense of the verb is important. “On Monday, Jack went to London”/ “On Monday Jack will go to London”, in these examples, Monday refers to different dates.

As dates can be fuzzy in natural language, they are represented in RDF as slice of time representing the incertitude about the event date. A date is represented by two dates that are equal in case of full date. For example, “2007” is represented by datebeg=20070101 and dateend=20071231.

Some dates are relative to national or religious dates. For example, “next Easter” needs to get the date for this year. This kind of event has dates changing each year like the Ramadan period.

Other relative dates depend on a date given previously or after in the text. Example: Massoud was murdered two days before the 9/11/2001.

6.3 Itineraries

Some texts can contain the description of an itinerary. Each sentence gives a part of the itinerary, but each sentence extraction gives partial information. Example: “John went to Istanbul on August 25 2004. Two days later, he arrived in Ankara. “

6.4 Some dates and/or places can be linked with following sentences

“John went to Lisbon on Monday. He met Jack. “ The extracted knowledge from the first sentence is quite complete. It lacks only the departure place which can be John’s residence place. But in the second sentence, the extraction gives only a meeting between John and Jack but no place and no date.

If no language construction prevents from understanding that it is the same place and the same date, the geochronolocalization information from the first sentence must be attributed to the second one. This kind of reasoning cannot be done without a study which is an ongoing process in our company. In what cases this inference can be done? With what succession of actions?

An examination of this problem on a corpus is done to find the situations where inference must be applied. **6.5 Processing of well known persons** For some well known people, anaphoras can be done using known particularities like the title of the person which is not linked in the document with the noun.

For example: “Sarkozy went to Marseille. After that, the president went to Nice.“

The fact that Sarkozy is the president is not mentioned in the document because it is well known during a time period. To solve this difficulty, databases on well known persons or organizations are used, like DBpedia. This action is under development. DBpedia files which have some errors are corrected manually before being used to tackle this problem of anaphora.

7. Evaluation

7.1 Inner evaluation

GEOLSemantics has developed its own evaluation system. This evaluation system is also used as non regression test between successive versions of the system. Texts are processed by the last successive versions. The RDF result is splitted into triples. Each triple is manually controlled if it has not been accepted or rejected by a previous version. Triples which are lacking are added by the evaluator. They are not added in RDF which is too heavy but are just reported by a message giving the waited information. When a new version gives the right triple, the text one is suppressed. With this tool, progression in output quality can be continuously estimated.

7.2 Evaluation campaign

The closer evaluation campaign is WEPS. There are no more campaigns organized but WEPS-3 data is available to test our system. Unfortunately, WEPS is only about attributes on persons without any attempt to extract actions.

We intend to test our system against WEPS-3 Data [4] by the end of this year. Our approach will be the extraction of all the relations we are able to extract and their use to perform a better clustering. Some extracted attributes are incompatible, for example birth place or date. Persons having incompatible attributes are considered as different persons.

In the case where there is no incompatibility, the proximity between persons will be calculated using the other attributes and the vocabulary contained in each documents. Vocabularies concerning a singer, a football player or a medicine doctor are strongly different.

The similarity matrix necessary for a clustering can benefit from the combination of the incompatibilities and the vocabulary contained in the texts.

The examination of the WEPS-2 data shows that there are different types of documents that should be processed in different ways.

For example, some documents are lists of homonyms. For each homonym, there is a list of id discriminative information like birth date, birth place, telephone number, address, etc. These documents must be splitted into subdocuments for each different homonym.

8. Conclusion

This work is an ongoing process. Information extraction in the French language is quite finished. English will be fully developed by August, Arabic and Chinese by the end of the year. This information extraction is included into the full SAIMSI process based on the WEBLAB platform from Cassidian. Application for semiautomatic filling of Ibase (I2) is possible.

9. Acknowledgements

This work has been done by the support of the Agence Nationale de la Recherche, Project SAIMSI (ANR-9CSOSG-08-01).

10. References

[1] Atkinson M., Piskorski J., Frontex Real-time News Event Extraction framework, *conference KDD’11*, August 21-24, 2011, San Diego, California, USA

[2] Fluhr C., SAIMSI, Suivi Adaptatif Interlingue et Multisource des informations, *workshop WISG’12*, 24-25 janvier 2012, Troyes, France

[3] I2 Limited, TextChart AutoMark 8, 2010 [4] Martin N., Khelif K., Focussed crawling using name disambiguation on search engine results, *International Symposium on Open Source Intelligence & Web Mining*, 12-14 September 2011, Athens, Greece

[5] Tanev H., Piskorski J., Atkinson M., Real-time News Event Extraction for Global Crisis Monitoring, *conference NLDB’08, processing of the 13th International Conference on Natural Language and Information Systems*, Springer Verlag Berlin, 2008

[6] Yangarber R., Jokipii L., Rauramo A. and Huttunen S., Extracting information about Outbreaks of infectious Epidemics, In Proceedings of the HLTEMNLP 2005, Vancouver, Canada, 2005.