SAIMSI

Suivi Adaptatif Interlingue et MultiSources des Informations

Christian Fluhr[1] and the SAIMSI contributors[2]

[1]GEOLSemantics, 32 rue Brancion, 75015 Paris

[2]See the full list at the paper's end

Christian.fluhr@geolsemantics.com

Résumé – Le but du projet est de suivre les activités de personnes suspectées d'activités illégales comme le terrorisme, le trafic de drogues, le blanchiment d'argent. Cette extraction est réalisée en français, anglais, arabe et chinois. L'extraction d'information est basée sur une analyse morphosyntaxique profonde. Elle reconnaît les mots simples, les expressions idiomatiques, les mots composés. Les relations syntaxiques de dépendance sont construites, les formes passives et actives sont identifiées, la négation et les modalités, la référence des pronoms et le traitement des temps composés des verbes sont réalisés. L'extraction d'information est propre à l'application et utilise des règles d'extraction sémantiques. A ce niveau certaines catégories d'entités nommées peuvent être changées. Cette extraction est basée sur une large ontologie de la sécurité. Le RDF produit alimente une base de connaissance. Celle-ci est munie d'un raisonneur qui infère de nouvelles relations à partir de celles issues des textes. Un module permet d'identifier l'auteur d'un texte à partir de l'apprentissage de signatures représentatives basées sur des critères morphologiques et linguistiques. L'interrogation de cette base permet de produire des visions des connaissances produites sous forme de fiches biographiques, de cartes géographiques, de frises chronologiques et de graphes de relations entre personnes et/ou organismes.

Abstract – The aim of the project is to follow activities of persons suspected of illegal actions like terrorism, drug traffic or money laundering. 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 large security ontology. The RDF obtained from the information extraction feeds a Knowledge Base. This Knowledge Base uses reasoning to infer new knowledge from the one obtained from the texts. A module allows recognition of the author of the text from learned signatures based on morphologic and linguistic features. Interrogation of the Knowledge Base can produce biographic sheets, geographical maps, timelines and graphs of person – person/organization relations.

1. Project's objectives

The aim of the project is to follow activities of persons suspected of illegal actions like terrorism, drug trafficking or money laundering from open sources of the Internet.

All open sources are concerned such as web sites, news, social networks.

Two media, text and speech are processed.

4 different languages French, English, Arabic and Chinese (Mandarin) are processed.

The gathered information is purified to extract only relevant natural language information; the language and coding are identified.

Information extraction is performed in each language but the relevant knowledge is represented in only one language (English).

In addition recognition of the author of a text or a speech is done according to signatures obtained by a learning on a representative corpus for each person.

Two different databases are built.

- A Knowledge Base which contains the extracted knowledge and inferred knowledge using inference rules.
- A cross-language fulltext database which contains the documents in their original language but which can be interrogated using natural language queries in only one language.

The Knowledge Base can provide a global view of the gathered documents according to the events and entity descriptions defined in the ontology. The Knowledge Base can be interrogated with complex queries. The results can be displayed using biographic

sheets of persons, geographic maps of events, timelines of events and graphical networks of person/person or person/organization relations.

The text database allows the user to retrieve information on themes that are not structured into the Knowledge Base.

When displaying a text, a user can ask for information on entities that are defined in the Knowledge Base.

On the other hand, a user can verify the content of the Knowledge Base by obtaining the documents that the knowledge has been extracted from.

The prototype is developed using Cassidian's WebLab® platform. This platform allows the different modules developed by GEOLsemantics, Mondeca, Cassidian and the LIP6 to interoperate. The IREENAT helps the consortium on the legal and deontological aspects that are particularly important in this project.

2. The SAIMSI Challenges

The scientific objectives of SAIMSI were very ambitious.

- Information extraction from texts in 4 languages based on a general purpose deep morphosyntactic analysis and a semantics application-dependent extraction.
- Creation of a large ontology of the security that satisfies both a top down vision of the user needs and a bottom up vision of the knowledge that can possibly be extracted from texts.
- Synthetic vision of the knowledge whatever the source text language.
- Identification of persons even when they are homonyms or have names with variation of spelling
- Processing of information in a flow (cumulative information)
- Recognition of a text's author

3. Legal and deontological aspects

Building database on persons is a very sensitive question concerning the protection of the citizen. It is a debating subject between the protection of private life and the protection of citizen against aggressions like terrorism, robbery, drug, etc. The law on informatics and liberty provides a framework allowing one to know if a database and its use are permitted.

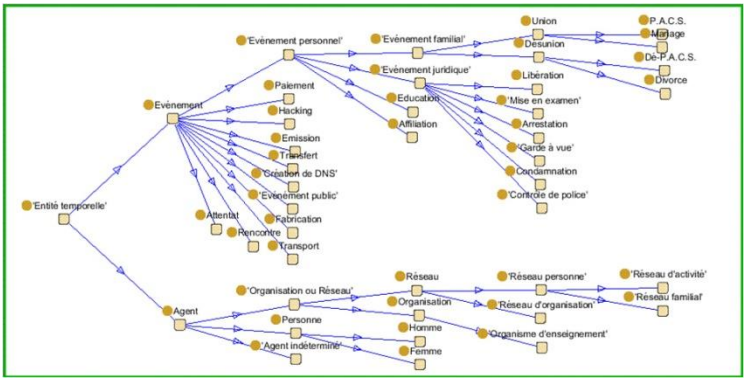
Difficulties arise because**,** for example, a tool like SAIMSI could be used by the police in legal conditions but as the consortium members are neither the police nor subcontractors of the police, they do not have the same rights and possibilities of processing this information.

There were a lot of interactions between the consortium and the CNIL. This has produced a better common understanding of this delicate matter.

This has resulted in a lot of restrictions concerning our experimentations: limitation of the search to 2 nouns concerning jihadism, anonymization of the other nouns, only one use case and destruction of the databases after the project has been finished.

4. Ontology of the security

The SAIMSI ontology is designed to allow us to express the information extracted from texts in order to transform them to intelligence knowledge on persons, organisations, locations and in general all named entities (brand names, events, etc.). The persons are described with biography identifier (name, birth date, address, nicknames, education, email, website, etc.),. The organizations include the CEO, headquarters, Growth rate, etc. The events mainly describe: meeting, transfer, marital status (marriage, divorce, etc.), juridical status (conviction, releasing, police custody, etc.), message delivery (interview, preach, fatwa, tweet, etc.), object manufacturing, payment, and many other events.



5. Information extraction

The information extraction developed by GEOLSemantics is done in 3 phases.

The first one is a domain-independent deep morphosyntatic analysis. Its role is to identify words, to recognize and categorize named entities, to identify syntactic (dependency) relations inside the noun and verb phrases and relations between the action and its dependents.

This morphosyntactic analysis is qualified as "deep" due to the negation analysis, verb tenses and modalities recognition, identification of pronouns referents and the active and passive forms identification.

The processing of pronouns is very important because if the referent of the pronoun is not identified, few useful information can be extracted. For example, in the

sentence “John goes to Paris with his brother Jack”, if the word “his” is not analysed as referring to John it will not be possible in the extraction phase to produce that John and Jack are linked by the family link “brotherhood”. It is also particularly important in processing biography where a lot of pronouns refer to the subject of the biography.

Negation and verb tenses and modalities are also very important from the semantics point of view. An action which is denied is of course completely different to an affirmative one. An action which is given in the future is only a potential action that must be verified at its fulfilment date.

Nicolas Sarkozy ira à Washington. »

Result :

Nicolas/first name
Sarkozy/family name
aller/verb future tense
à/preposition
Washington/proper noun

Named entity of person :

Nicolas/first name, Sarkozy/proper noun

Named entity place :

Washington/proper noun

Subject-verb:

Subject : Sarkozy Verb : aller

Verb-complement relation:

Verb : aller Complement : Washington

The second phase is the semantic extraction. This step is largely facilitated because the result of the deep morphosyntactic analysis simplifies the writing of the rules.

At this level, the named entity category can be modified if the extraction rule considers that the role of the entity is incompatible with its previous category.

For example in the sentence “Paris declares “….”

At the morphosyntactic level Paris is considered as a place. But in the context of the action of emission of a message, the agent cannot be a place. In fact, the emission can only be done by a person or an organization and if at the origin it was considered as the capital of France, the new category is an organization and one can infer that it is the French Government.

A series of extraction rules are written according to the actions and attributes of persons that are described in the ontology.

The result is expressed in RDF.

```
<gs:Sent rdf:nodeID="0">
  <gs:hasTriple>
    <gs:Transfer rdf:nodeID="id16Transfer">
<gs:authorValidation>future</gs:authorValidatio
n>
      <wn:undergoer rdf:nodeID="id8Person" />
      <gs:locend rdf:nodeID="id22Location" />
    </gs:Transfer>
  </gs:hasTriple>
  <gs:hasTriple>
    <v:Location rdf:nodeID="id22Location">
      <gs:location-
name>Washington</gs:location-name>
    </v:Location>
  </gs:hasTriple>
  <gs:hasTriple>
    <foaf:Person rdf:nodeID="id8Person">
      <v:n>
        <v:Name>
          <v:given-name>Nicolas</v:given-name>
          <v:family-name>Sarkozy</v:family-
name>
        </v:Name>
      </v:n>
      <foaf:gender>male</foaf:gender>
    <foaf:Person>
  </gs:hasTriple>
</gs:Sent>
```

Example of extraction for the sentence “Nicolas Sarkozy ira à Washington”

The 3rd phase was not really taken into account during the project preparation. If we consider the result of semantics extraction at the sentence level, even if pronouns are treated, a lot of information is not present as a human can infer them only by a sequential reading of the document.

This concerns several aspects:

- Processing of incomplete dates: there are two cases
  - The date is relative to the document production date. Ex: “*Monday* the president goes to Nantes.”. The verb tenses are also important ex: Monday the president *will* go to Nantes” don’t give the same date as the one of the preceding sentence.
  - The date is relative to a date previously given in the document. Ex: “Massoud was killed September 9, 2001.

Two days later the twin towers …

- Duplication of temporal and/or spatial information from a sentence to the following one. This is only applicable for some particular couple of actions like travel – travel or travel – meeting. Ex: John went Saturday to Istanbul. He will go then to Baghdad. In the second sentence the departure point is not given but the reader assumes that it is Istanbul because it is the arrival point of the previous sentence.
- Identification of an entity in different occurrences: in a text, persons are mentioned using different character string. For example: Barack Hussein Obama, Obama, President Obama, US president. Generally, the first occurrence gives more information. If several persons with the same family name are in the same document, the author gives discriminative information (first name, title) to distinguish them.

6. Knowledge base and reasonning

The philosophy of the Knowledge Base use(i.e., a data base driven by an ontology-driven domain data model) into the SAIMSI platform is to bridge the gap between the information extraction content annotation process, and the knowledge repository storage. In so doing, we have set up a middleware which has the purpose to handle the information extraction results and to populate an ontology-driven knowledge base with the extracted annotations.

To achieve this goal, this middleware called CA-Manager [8] relies on the recommendations made by the W3 Consortium and the Semantic Web community:

1. Express the knowledge using RDF[1]/OWL[2] languages;
2. Set up a service-oriented architecture (SoA) to feed other systems and display rich results.

CA-Manager is composed of 5 main functional components that support building and managing customized workflows for semantic contents annotations, ontology population and ontology-based information extraction systems:

1. Extraction: extract knowledge from content;
2. Consolidation: reconcile extracted knowledge with the domain ontology and the content of the knowledge repository;
3. Storage: export and store the reconciled knowledge;
4. Validation: let the human user validate the suggested annotations and knowledge;
5. Enrichment: export term and entity candidates into the information extraction linguistic resources.

Once the knowledge base is set up, one can perform reasoning tasks. An inference engine is used upon the knowledge base in order to achieve closure of the semantic graph. The reasoning rules conform to the ontology-model (domain/co-domain, restrictions, cardinalities, functional properties) are performed to meet business needs (e.g., to deal with homonym disambiguation). This knowledge base closure is applied in order to validate constraints and infer new knowledge. The use of the inference engine relies on the following types of reasoning:

- Logical reasoning: create relations between the candidates (matching or distinction);
- Candidate disambiguation: for example, two nicknames that point to persons in two different places at the same time necessarily imply two different persons.

We performed a number of inference rules in order to perform the knowledge base closure.

7. Identification of the author of a text

Identifying the author of a text is an important building block in this project. Unfortunately the authorship attribution literature demonstrates the difficulty of such a task. Actually it appears to be difficult to design a classifier that outperforms naïve strategies such as using a linear classifier operating on a number, e.g. the most frequent, of lexical features such as character trigrams or word counts. Yet such systems reach only limited accuracy. To overcome this difficulty we proposed to use a statistical method called feature bagging that relies on learning classifiers on different random subset of features, then to combine their decision by making them vote. It is called an ensemble method.

Many methods have been proposed for combining classifiers such as co-training, boosting, bagging, a number of which have been designed or adapted for working with classifier exploiting different subsets of features [1], [2]. In particular, feature bagging has been investigated by a few researchers in the past. Viola & Jones [1] used boosting with extremely weak classifiers (learned on a single feature each) every iteration. [3] also used boosting with an adaptation of AdaBoost to

[1] Resource Description Framework (http://www.w3.org/TR/rdf-primer/)

[2] Ontology Web Language (http://www.w3.org/TR/2004/REC-owl-features-20040210)

feature weighting instead of samples weighting as in AdaBoost.

In our work we decided to investigate a standard bagging combination where an eventually large number of base classifiers are learned on random subsets of the features (with eventual overlap) and are then combined at test time through a voting procedure. In practice we investigated using a majority vote decision process with a number of SVM classifier trained on many (hundreds to thousands) random subsets of few (tens to hundreds) features. SVM classifier are learned with the Libsvm toolbox. Experiments on a few corpus, including a participation to the PAN 2012 challenge at CLEF 2012 [4], have shown that this approach allows significantly outperforming the standard benchmark method exploiting a SVM working on all features together (see table below).

Model	Accuracy
Bagging with 600 classifiers using each 100 features	79.4
Bagging with 600 classifiers using each 225 features	76.7
Bagging with 600 classifiers using each 600 features	76.1
Naïve approach: SVM with all 3000 features	71.6

Performance of the Bagging feature approach on a blog corpus with 60 authors. Performance of the naïve approach is given for comparison.

8. Identification of persons

The identification of persons is crucial from the point of view of efficiency for the following up of illegal activities but also in order to prevent persons from being accused for illegal actions performed by homonyms.

It is a well known problem of synonymy and homonymy.

As usual the synonymy problem is easier to solve.

Processing of person name, synonymy could be necessary when the original spelling of a name was in a non-Latin character set like Arabic, Chinese or Russian.

Romanization of these nouns is not normalized and change from a country to another.

Another case is for nouns that are only known by their pronunciation (from wire tapping for example). Spelling variants must be considered as possible names of the person.

Example: for Oussama ben Laden, some variant produced by SAIMSI transliteration tool [6] and found in the Internet.

osama bin ledan
osama bin Laden
ozama ban ladin
osama ben ladane
osama ben ladin
asama binladdan
osama binleden
osama bin lden
osama ben ledan
usama bin laddan
…..

The homonymy problem is much more complicated. The consortium has only begun the work on this problem because the discussions with the CNIL about the possibility of producing and use a database of persons for homonym recognition was very long.

The first results obtained in English using the data of the WEPS-3 campaign [5] give encouraging perspectives.

Discriminating on one first name last name couple "Alan Cox" on 200 pages in English manually controlled, shows that discriminating homonyms using information extraction developed for SAIMSI gives good results.

The algorithm to clusterize the documents is based on:

- a set of attributes of incompatibility. If birth dates or birth places are incompatible, the persons are different
- a set of absolute compatibility. If two telephone numbers are compatible (same local number) it is the same person.

For the other cases, a statistical distance based on non absolute criteria and vocabulary is used.

The results show that in the 200 pages there were 19 different persons. 13 of them are represented in only one document. 75 documents do not contain any information from the SAIMSI ontology. These documents have been attributed to the right person using the vocabulary. We were probably lucky because all the 19 persons have very different activities except Alan Cox one of the fathers of Linux and a professor in computer sciences.

9. Multilinguality

One of the important aspects of this project is the management of multilinguality. This aspect is particularly hard because the chosen languages are very different from the linguistic and cultural point of view like French, Arabic and Chinese.

the consequence is an obligation to design the morphosyntactic analysis as general as possible to be able to manage a maximum of languages phenomena.

The cross language text database follows the principles elaborated originally by the EMIR European project. Documents are indexed in the source language. Interrogation is done using bilingual reformulations. The interrogation can be done in the user's mother tongue and the results are displayed in the original language

guarantying that the information has not been alliterated. A machine translation system can be used to have a rough idea of the text content. The crosslanguage interrogation can be considered as a good tool for the decision of human translation because it shows that a document is relevant for a particular question.

The multilingual management of the knowledge base is different. To be able to merge information coming from a text in different languages the extracted knowledge is coded in only one, which is English.

Producing a representation in only one language brings problem similar to machine translation. The problem is simplified by the fact that the extraction is done in semantic domain which is strongly limited by the ontology.

It is sometimes necessary to produce very large authority lists like the list of jobs in each language.

Even in the limited semantic domain, ambiguities can occur to get the right translation. This happens in the case a language has a single word to express 2 concepts that are distinguished in English. It is the case for "belle fille" in French which can be in English : « **daughter-in-law »** and **« stepdaughter ».**

If it seems to be difficult to resolve the ambiguity using the extraction rules, an alignment on the most ambiguous language is done.

The last problem occurs by the translation of named entities in the context of character set change. The problem is particularly important in our project where 3 very different character sets are used: Latin, Arabic and Chinese.

For well known entities, bilingual dictionaries can be used but it is impossible to use this approach for all the possible person nouns or place nouns that can be encountered in a text.

Among the possible romanizations we choose one to translate the unknown name.

For Chinese we have chosen the pin-yin without accents.

The name of the previous President 锦涛**胡**

Is represented in pinyin by Hú Jǐntāo. Can be simplified without accents by Hu Jintao

For the Arabic we have chosen the ALA-LC Romanization Table

For exemple :the name of the Algerian President "بُوتَفْلِيقَة الْعَزِيز عَبد

is represented by:

ʻabd al-ʻazīz būtaflīqah **(Abdelaziz Bouteflika)**

To compare with names written in French or English we use the previously described transliteration tool which produce all compatible spelling variants.

10. Crosslanguage interrogation of the text database

The crosslanguage interrogation text database follows the concepts developed during the European project EMIR.

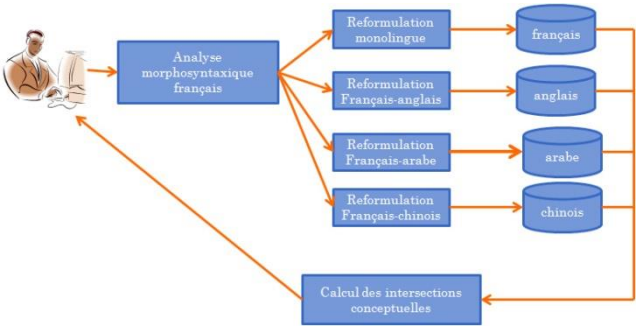
In SAIMSI, the system is built upon the open source Lemur/INDRI text database.

The general principle is the same level of morphosyntactic processing for text in document source language and queries.

The result of the morphosyntactic processing of the query is processed by the reformulation tool that infers equivalent concepts in the same language for monolingual interrogation and translations for crosslanguage interrogation.

A comparator elaborates the best concept intersection between the query and the document whatever the document source language.

In case of precise queries the translation ambiguities are resolved in the relevant documents.



Cross-language interrogation of documents in 4 languages

The results are presented in a list of classes sorted by a decreasing order of relevance. Each class is characterized by the concept intersection.

For example for the query "meurtre de Massoud" the best class is characterized by "meurtre-Massoud" that represents the syntactic relationship between the concept of murder and the person who was killed.

The documents in this class contain for English documents "assassination of Massoud", "murder of Massoud", 'Massoud's murder".

The document viewer displays the relevant part of documents with named entities displayed in different colors and the words representing the query concepts are highlighted.

11. Components integration

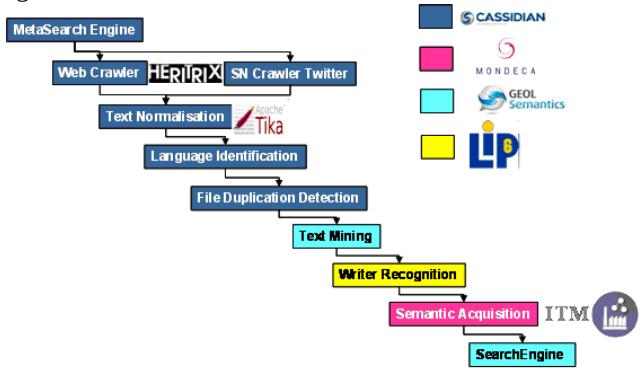
For each of the objectives previously described, software components were prototyped in order to design solutions that would effectively address the specific concerns. For a number of technical reasons, the developments made by the project partners used various technologies. These components had not been designed

to work together and had little capability to communicate with each other. However, thanks to the WebLab platform, these heterogeneous components were smoothly incorporated into a consistent processing chain and were able to interoperate. In this way, a comprehensive application could be implemented to make the most of complementarily of the components and to enable demonstrations of capabilities regarding the operational need.

WebLab is an open source framework developed and maintained within the OW2 Consortium (http://weblab-project.org) since 2009. It allows to expose the native functions of existing components as services and to assemble these services within a processing chain. To achieve this, WebLab proposes a conceptual information model to define a common exchange format and facilitate the orchestration of the processing services: a producer service encodes its results according to this exchange format and provides them to a consumer service, which will decode the data and then process them. The orchestration is thus rationalized since it does not involve specific interfaces between each service. The use of unique data format also reduces the computational and development costs and the introduction of new services is facilitated.

The OW2 forge hosts several existing services based on open-source components. These services are fully compliant with the WebLab exchange model and can be easily reused in various applications. The SAIMSI project reused some of them to collect documents, to normalize their content, to identify their languages, to eliminate duplicate, etc.

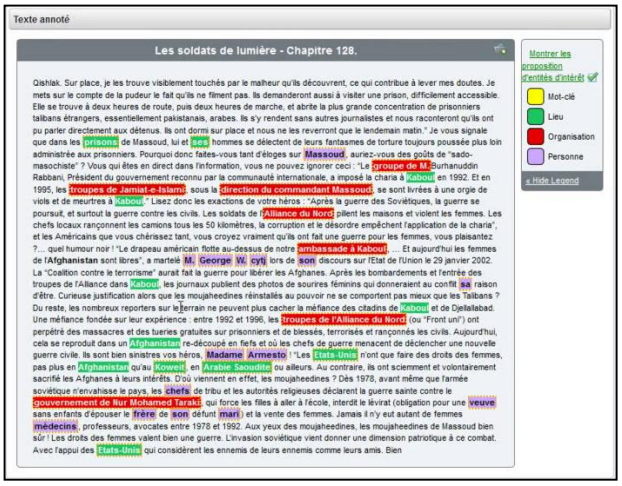
The SAIMSI processing chain is represented in the figure below:



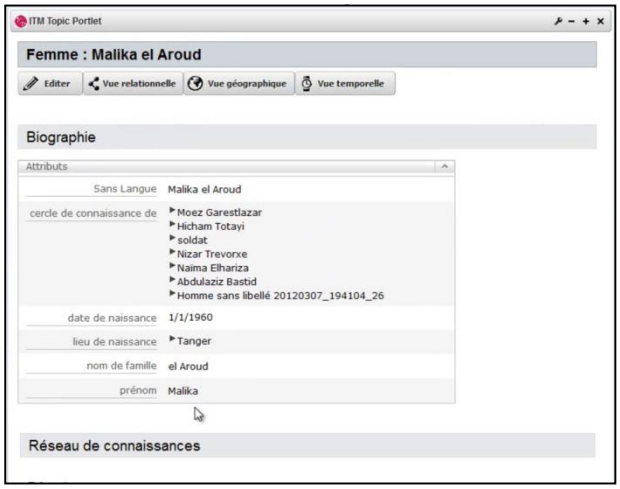
Processing chain of the prototype

The modularity aspect has also been a keypoint for the Graphical User Interface components. All GUI components were integrated into the WebLab Portal which is based on the "Portlet" technology and the free and open source enterprise portal Liferay. In the same way as for service integration, these choices promote reuse and composition of components to develop specific application. They also enable “on-the-fly” GUI composition by the user. Some new portlets were developed to cover the SAIMSI needs but some others had been produced in previous projects and were available.

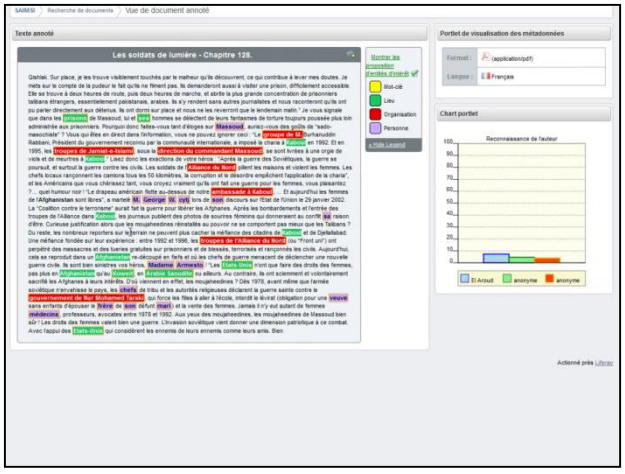
Some examples of portlets and composed pages are presented below:



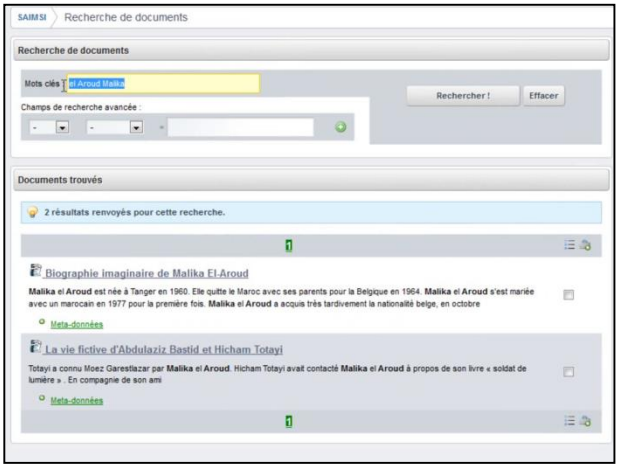
Portlet to display the information extraction



Portlet to display the Knowledge Base content



Composed GUI to display the author of a text



Composed GUI to inquire the text database

12. Conclusion and Perspectives

The aims of the SAIMSI project were very ambitious. Each challenge could be a full project. The operational subject of SAIMSI brought a lot of legal constraints preventing the consortium to experiment the system on real situations.

It remains a lot of work to industrialize the full system.

Some ameliorations have been identified and should be the objective of new researches:

- Identification of the document type. The SAIMSI system is well tuned for news articles but particular processing must be done for bibliographies, CVs, biographies and dialogues in forums.
- The identification of the text creation date is very important for relative dates processing. There is no normalization to identify this date and a parser must be developed for each site.
- Extension to SMS and incorrect languages and Arabic Dialects written in Arabic or with latin characters.

Processing time, especially in the information extraction must be ameliorated for mass processing. This work has begun. Semantic extraction time has been divided by 100. Industrialization of the morphosyntactic analysis is on the way and will be available in April 2013. We expect to divide the time by 500.

Even if the full prototype is not ready to be used in professional use, some of the modules can be used within a short time in real applications. It is the case for the information extraction if used as a productivity tool for the introduction of information from text. A good example is the processing of charge sheets to fill structured databases like I2 of IBM. Today the I2 system is filled manually from the charge sheets. A processing of charge sheets with the SAIMSI information extraction tool with a control of results by the investigating officer will sharply decrease the input time. An extension of the semantic representation has been done by GEOLSemantics and a first tool for control and modification of the extracted information has been created.

Références

[1] P. Viola, M. Jones: *Rapid object detection using a boosted cascade of simple features.* In: Computer Vision and Pattern Recognition, 2001

[2] C. Sutton, M. Sindelar, A. Mccallum: *Feature bagging: Preventing weight undertraining in structured discriminative learning*. Tech. rep., CIIR (2005)

[3] J. O'Sullivan, J. Langford, R. Caruana, A. Blum,: *Featureboost: A meta learning algorithm that improves model robustness*. In: In Proceedings of the Seventeenth International Conference on Machine Learning. pp. 703–710 (2000).

[4] F-M. Giraud et T. Artières, *Feature Bagging for Author Attribution*, CLEF 2012.

[5] J. Artiles, A. Borthwick, J. Gonzalo, S. Sekine, and E. Amig_o, *WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks*, third WePS evaluation workshop, 23 sept 2010, Padua, Italy.

[6] H. Saadane, A. Rossi, C. Fluhr, M. Guidere, *Transcription des noms arabes en écriture latine*, SETIT 2011, , March 23-26, 2011 – TUNISIA

[7] C. Fluhr, A. Rossi, L. Boucheseche, F. Kerdjoudj, *Extraction of information on activities of persons suspected of illegal activities from web open sources*, conference LREC2012,workshop "language resources for public security applications", 27 may 2012, Istanbul, Turkey.

[8] F. Amardeilh. *Semantic annotation and ontology population*. In J. Cardoso and M. Lytras, editors, Semantic Web Engineering in the Knowledge Society. Idea Publishing, 2008.

Acknowledgement

This work (SAIMSI Project) has been funded by the ANR CSOSG program (reference **ANR-09-SECU-08-01)**

Co-Authors :

Patrick Giroux, Cassidian, patrick.giroux@cassidian.com
Emilien Bondu, Cassidian, Emilien.Bondu@cassidian.com
Thierry Artière, LIP6, thierry.artieres@lip6.fr
François-Marie Giraud, LIP6, giraudf@poleia.lip6.fr
Hacene Cherfi, Mondeca, hacene.cherfi@mondeca.com
Florence Amardeilh, Mondeca, florence.amardeilh@mondeca.com
Halima Dahmani, GEOLSemantics, halima.dahmani@gmail.com
Aurélie Rossi,GEOLSemantics, aurelie.rossi@geolsemantics.com
Louise Boucheseche, GEOLSemantics, louise.boucheseche@geolsemantics.com