An Infrastructure for the Semantic Annotation and Mining of Documents in the Public Security Domain

Francesca Frontini, Carlo Aliprandi, Clara Bacciu, Roberto Bartolini, Andrea Marchetti, Enrico Parenti, Fulvio Piccinonno and Tiziana Soru
Synthema, Genesy, ILC CNR, IIT CNR - PISA - Italy
Public security domain

**GLOSS** is an integrated system for the analysis and retrieval of data in the environmental and public security domain.

Efficient access to information is crucial in the work of organizations that require decision making in emergency situations (such as natural disasters).
Public security domain

Users from the Civic Defence Agencies need to deal with large quantities of information often in a short amount of time. Their source of information lies in internal and public documents reporting previous environmental issues for the relevant areas.

Historical reports can be used:

● to predict environmental risk for certain area;
● in case of emergency in a given area, to efficiently extract information about previous events.
Public security documents

FEMA  www.fema.gov

Protezione civile  www.ilgiornaledellaprotezionecivile.it

Currently GLOSS has **crawled** a document base of
**480** public documents in English
**1200** in Italian
The project

We shall briefly present the GLOSS infrastructure and its use, and how semantic information of various kinds is integrated, annotated and made available to the final users in order to better retrieve information from texts.

GLOSS is a regional funded project (Tuscany - Italy): final users from local civic protection agency.

Previous experience of KYOTO (ICT-211423).
Goal

- allow for an enhanced and multimodal search of topics of interest for the community;

- further enrich the available resources by adding new domain concepts to the knowledge base (WordNet) and new geographic information (e.g. new locations, polygonal areas) to the Geodatabase (Geonames).
Mercoledì 18 Gennaio 2012. Sono 10 i feriti a seguito della violenta esplosione di un metanodotto avvenuta oggi nella frazione di Lupinaia in Garfagnana (Lucca)

[Wednesday 18th of January 2012. 10 people were wounded in a violent explosion of a methan pipeline today in the village of Lupinaia in Garfagnana (Lucca)]
GLOSS - infrastructure
GLOSS knowledge extraction

Candidate terms extraction (dependencies & Log Likelihood)

**minaccia**  [Minaccia, minacce, minaccia]  **ROOT**  minaccia  N  267  
minaccia<>radiologico  
minaccia<>chimico  
minaccia<>biologico  
minaccia<>nucleare  
minaccia<>proveniente  
minaccia<>spaziale  
.....
GLOSS knowledge extraction

Human intervention

**minaccia** [Minaccia, minacce, minaccia]  ROOT  minaccia  N  267

- minaccia<>radiologico
- minaccia<>chimico
- minaccia<>biologico
- minaccia<>nucleare
- minaccia<>proveniente
- minaccia<>spaziale

...
GLOSS geo-knowledge extraction

automatically extract new candidate terms (NER + heuristics)
"frazione di Lupinaia, in Garfagnana"

manual intervention
Lupinaia

geobase

- **name**: Lupinaia
- **coordinates**: 44.126502 - 10.401306 + area
- **is-in**: Provincia di Lucca
- **is-a**: populated place
Gloss Geo editor
GLOSS - semantic search

Why should users be willing to make the effort?
Run queries such as:

"find all catastrophic events occurred in the selected time and area"
Mercoledì 18 Gennaio 2012. Sono 10 i feriti a seguito della violenta esplosione di un metanodotto avvenuta oggi nella frazione di Lupinaia in Garfagnana (Lucca)
Exploiting converging interests

Organizations who are dealing with large document bases are offered a tool to manage and search their documents in a more efficient way and or to make them more easily accessible by third parties.

LT experts can meet this need by providing tools for semi automatic document treatment while at the same time gathering domain knowledge by experts.
Future work

Digitalization:

... a lot more documents out there to be made effectively accessible to their owners

and

... exploited for text mining purposes!
Conclusion

We described GLOSS, an architecture for the analysis and semantic mining of document in the public safety domain. This system, combining state of the art natural language and semantic processing, a refined geographic data management system and methods for semi-automatic collaborative definition of knowledge by experts, can be used by organizations in the public safety domain to efficiently organize, increment and access their domain knowledge, which in turn can be crucial to collect the necessary information for decision taking in emergency situations.