Semantic Press

Eugenio Picchi, Eva Sassolini, Sebastiana Cucurullo, Francesca Bertagna, Paola Baroni
Consiglio Nazionale delle Ricerche – Istituto di Linguistica Computazionale (CNR-ILC)
Via Giuseppe Moruzzi N° 1 – 56124 Pisa – ITALY
{eugenio.picchi|eva.sassolini|nella.cucurullo|francesca.bertagna|paola.baroni}@ilc.cnr.it

Abstract

In this paper *Semantic Press*, a tool for the automatic press review, is introduced. It is based on Text Mining technologies and is tailored to meet the needs of the eGovernment and eParticipation communities. First, a general description of the application demands emerging from the eParticipation and eGovernment sectors is offered. Then, an introduction to the framework of the automatic analysis and classification of newspaper content is provided, together with a description of the technologies underlying it.

1. The Framework of *Semantic Press*: the *Linguistic Miner*

The *Semantic Press* (SP) tool was born as an evolution of a complex system, the so-called *Linguistic Miner* (LM, Picchi et al., 2004), set up in 2003 with the aim of developing a framework for the automatic extraction of linguistic knowledge from very large amounts of texts (from different sources and in different formats) to be exploited in didactic, editorial and cultural products. SP uses language resources extracted from LM, but adopts a lot of the distinguishing operating modalities and analysis tools of this system as well.

Building LM involved two fundamental steps: firstly, the data were gathered; secondly, they were linguistically analysed to be further processed and classified. The first step produced a repository (a "mine") of around 200 millions words, together with an automatic topic classification of texts. This was achieved by exploiting procedures for the upgrade and increase of textual data within the "mine" and for the automatic acquisition of data from the Web through the Spider technology, both with periodic updating and by means of user-defined paths. The "mine" is thus constantly enlarged in size. The second step consists in the automatic linguistic processing of the textual material collected by using the modules of the PiSystem (Picchi, 1994), an integrated framework for the processing of textual and lexical material, whose most important module is the so-called *Data Base Testuale* (Textual Data Base, in short DBT). The most effective procedures for further analyses of texts are POS tagging and lemmatization, which were performed on the whole repository.

The linguistic analyses performed and their respective annotations were also recorded within the "mine", which is therefore not only a repository of textual materials, but also a database linguistically annotated.

Text Mining techniques are applied and exploited in a lot of frameworks: for instance, Inxight's LinguistX (Inxight White Paper, 2006), IBM's Intelligent Miner (Intelligent Miner White Paper, v.8.2), TextWise1 etc. In this scenario, LM stands out for its being based on tools and basic technologies developed to carry out good linguistic analyses supporting language resources for news applications. LM is specifically tailored for analysing Italian, but is obviously open to other languages.

In the last year, LM was addressed to meet the needs of political and institutional bodies (such as Regione Toscana), which expressed their interest to use and exploit a tool for an intelligent access to the flow of news and information provided by Italian newspapers available on the Internet. This aim is in line with the current interest of CNR-ILC in the topics of eParticipation and eGovernment, also confirmed by its participation in the DEMO-net project (http://www.demo-net.org/).

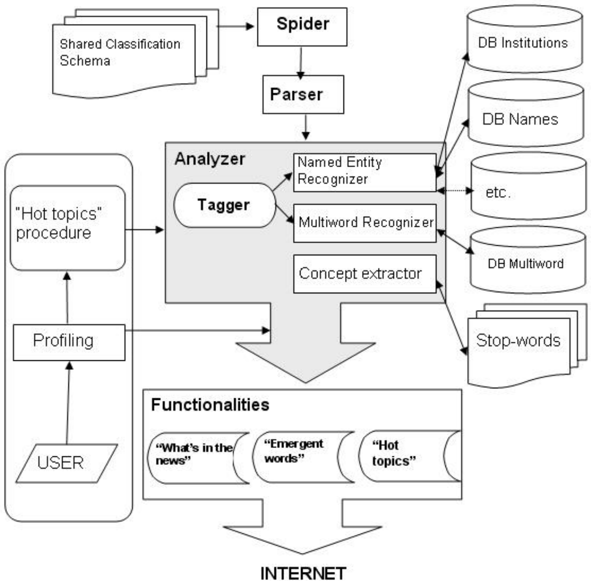


Figure 1: the SP system schema.

2. Application Scenario and Technological Challenges

eParticipation is the extension and transformation of citizens' participation in political deliberation and decision-making processes through Information and Communication Technologies (ICT). The notion is complementary to the eGovernment one, which concerns more the use of ICT to improve and innovate the quality of services offered to citizens by the Public Authority.

A very significant issue for eParticipation and eGovernment is the necessity – for citizens, but for professional politicians and consent-making operators as well – to know salient facts and features, hidden amidst a very large quantity of data, which stand out for their frequency: that allows to get interesting and constantly updated information about trends, tendencies and major topics in a given period. For this kind of needs, fed by the availability of a huge quantity of information, constantly changing and distributed in a large number of Web sites, Text Mining techniques can be very useful and full of promise.

3. Semantic Press

SP customizes some of the functionalities of LM to the analysis of the information available in Italian on-line newspapers. A demo offering an overview of the tools used can be found at the following URL: http://serverdbt.ilc.cnr.it/edicola/SemanticPress.htm. SP is different from other tools for the automatic press review (such as, for instance, PressToday, findable at the URL http://test.presstoday.com): as a matter of fact, it is not only a way to present and incrementally store news and articles pertinent to different sectors, but also a powerful tool, based on NLP and Text Mining techniques, for highlighting emerging subjects, issues and words and for adapting analyses of news to domains of interest and individual users.

Every morning, SP downloads all the articles of the most important and read Italian newspapers: Il Sole 24 Ore, La Stampa, La Repubblica, Il Corriere della Sera, Il Messaggero, La Nazione, Reuters, Il Tempo, Il Tirreno (including the local editions of Tuscany). The system follows a parametrical approach to increase the cardinality of the newspaper sources to be acquired. The acquisition procedure not only downloads the title and text of each article, but also visits and saves all the textual material available in the Web pages linked to the articles. Some filters are activated in order to avoid downloading dossiers, tables and other uninteresting sources. During the day, SP performs the updating of the articles downloaded by adding new information, if available, and handling cases of similarity between different versions of the same article. In this way, about 1200 new articles are stored every day. The textual material acquired is saved and converted in an internal format based on DBT specifications. When the article is saved, it is also classified according to a pre-defined set of ten topics (politics, finance, sports etc.). The classification is based on the classifying tags already present in the sources, which are normalized and mapped onto a shared and common classification scheme.

Then, the articles are linguistically analysed in order to obtain texts annotated at lemma and POS levels. Two basic technologies are used in the analysis phase: named entity recognition and multi-word recognition. Different experiments are carried out to evaluate both the impact and performance of the basic technologies exploited and the various strategies adopted.

4. System Functionalities

The main components of the SP system are illustrated in this paragraph. Subsequently, the innovative aspect distinguishing it from the other on-line press reviews will be evaluated. Finally, interesting perspectives for new developments and applications of the methodology implemented for SP will be proposed.

4.1 Emerging Words

By using lists of ad hoc stop-words able to filter insignificant words (articles, prepositions, verbal forms, days, months, numbers etc.), the system offers users the possibility of visualizing the words recurring more frequently in daily articles. The selection of the empty words to be inserted was made initially after an exam of the characteristics of the articles and subsequently after periodical checks on the validity of the results. That guarantees the correctness of the procedure, but calls for continuous updates and controls of data.

The aim is to make all the different words of an article emerge by providing a general overview of the predominant vocabulary in a particular day.

The concept of "emerging words" can be summarized with the assertion that it substantially provides two kinds of information: firstly, it immediately indicates the typologies of the most relevant news that are present in the individual genres every day; secondly, it provides indications about the terms mostly used in the various typologies of news.

In the case of sports, for instance, it can be noticed how football teams such as Inter and Milan are identified more frequently by the terms "neroazzurri" and "rossoneri", respectively, instead that by their names: an evidence, this one, of the influence of the football slang on the style of sports articles.

4.2 What Is in the News?

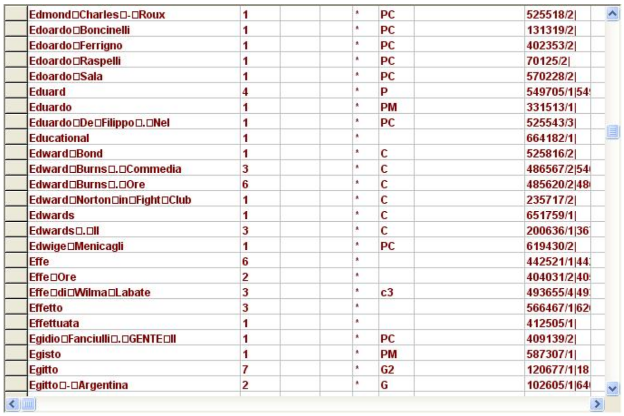
SP presents the most important themes emerged from the automatic analysis of daily news. Every day users can know the most discussed topics in the forms of names of persons (often politicians, but also show business people, important scientific or artistic personalities, protagonists of crime, of news etc.), locations of important happenings, facts, events etc.

This functionality, borrowed from the procedures of LM, is based on tools for linguistic analysis and language resources able to identify named entities and multi-words.

4.2.1 Named Entities

SP provides a list – ordered alphabetically and articulated by domain of interest – of the main named entities characterizing the pages of the daily newspapers; in order to do that, it makes use of modules and procedures peculiar to the LM project, further developed and articulated.

The component for the identification and annotation of named entities is a complex system and is based on the integration of different approaches. The tools inherited from LM are at the base of the system and their application to large quantities of textual material found in the Internet allowed the creation of great repositories of elements classified as certain or potential components of named entities. The main typologies of classified elements can be summarized in proper nouns, surnames, geographical names, business names and typical connectors for the composition of more complex named entities.



Edmond□Charles□-□Roux	1			*	PC		525518/2|
Edoardo□Boncinelli	1			*	PC		131319/2|
Edoardo□Ferrigno	1			*	PC		402353/2|
Edoardo□Raspelli	1			*	PC		70125/2|
Edoardo□Sala	1			*	PC		570228/2|
Eduard	4			*	P		549705/1|54
Eduardo	1			*	PM		331513/1|
Eduardo□De□Filippo□.□Nel	1			*	PC		525543/3|
Educational	1			*			664182/1|
Edward□Bond	1			*	C		525816/2|
Edward□Burns□.□Commedia	3			*	C		486567/2|54
Edward□Burns□.□Ore	6			*	C		485620/2|48
Edward□Norton□in□Fight□Club	1			*	C		235717/2|
Edwards	1			*	C		651759/1|
Edwards□.□Il	3			*	C		200636/1|36
Edwige□Menicagli	1			*	PC		619430/2|
Effe	6			*			442521/1|44
Effe□Ore	2			*			404031/2|40
Effe□di□Wilma□Labate	3			*	c3		493655/4|49
Effetto	3			*			566467/1|62
Effettuata	1			*			412505/1|
Egidio□Fanciulli□.□GENTE□Il	1			*	PC		409139/2|
Egisto	1			*	PM		587307/1|
Egitto	7			*	G2		120677/1|18
Egitto□-□Argentina	2			*	G		102605/1|64

Figure 2: Named Entity intermediate results.

By using a simple grammar, some rules were written for the acceptability of command strings that are potential candidates for the annotation as named entities. The integration between specialized resources and rules for recognition is the core of our system for the Named Entity Recognition (NER). This solution allowed the recognition and classification with a high percentage of success because a verification of the rules – and, if necessary, the creation of new ones – has always corresponded to the increase of the resources.

A particular attention was paid to the identification of named entities characterized by the presence of key words (headers) of great importance and productivity, such as universities, institutes, academies, churches, museums etc. In order to solve such phenomena with a high degree of reliability, a procedure was realized, which automatically searches all the possible candidates in the Internet for each header identified. The results of the procedure were stored in a special database used in relation to the database of named entities. In order to provide an evaluation of the performance of the NER component, a generic day was used as a sample, the automatic system for classification was applied, the percentage of recognized named entities was measured manually and, in the same way, errors and information gaps were analysed.

The results of the procedure are reported below.

Test carried out over 1200 articles estracted from 10 on-line headings (La Repubblica, Il Corriere della Sera, La Nazione, Il Tempo, Il Tirreno, Il Sole 24 Ore, Il Messaggero, Reuters, La Stampa, Lettera 22) on 27th March 2008.

Total number of forms: 33144.
Total number of occurrences: 535652.
True Positive = 22435.
True Negative = 795.
False Positive = 470.
False Negative = 2682.

$$\pi = \frac{TP}{TP+FP} = \frac{22435}{22435+668} = 0.9710$$

$$\rho = \frac{TP}{TP+FN} = \frac{22435}{22435+2682} = 0.8933$$

$$F_{\beta=1} = \frac{2*\pi*\rho}{\pi+\rho} = \frac{2*0.9710*0.8933}{0.9710+0.8933} = 0.9305$$

Task	Precision (π)	Recall (ρ)	F-measure
NER	97%	89%	93%

Table 1: results of the NER system.

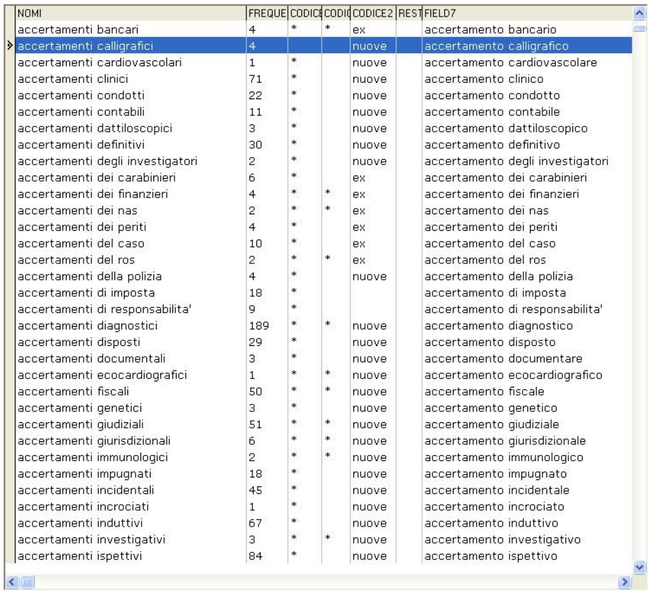
From the results obtained it can be deduced that the tool has a high degree of precision and a good degree of recall. The difficulty in improving the NER system is due to the fact that the domain taken into account is not a specific domain where it is possible first to train it and then to recognize terms with a high degree of recall. The typologies of textual material processed by the system are several, ranging from sports to news, from foreign affairs to culture. In articles, in fact, new proper names, geographical resorts, abbreviations etc. appear more and more frequently, which calls for a periodical update of the databases and of the rules of extraction.

The update procedure is computer-based within the system in order to periodically suggest the novelties to be inserted in the database interactively.

For completeness of information it can be added that the system is equipped with other two components: one identifying addresses and another identifying titles such as “Signor Tizio” (“Mr Tom”), “Presidente della Repubblica Sempronio” (“President of the Republic Harry”) and similar ones. Such components are not described here since, at the current state of the development of the system, they are not considered useful for a further exploitation of SP.

4.2.2 Multi-words

Even the component for the recognition of multi-words arises from the LM system. The procedure is based on the analysis of large quantities of texts inherited from LM and of the linguistic components forming them. In a first phase, texts are analysed by means of procedures of morphosyntactic classification. Subsequently, a software module applies rules for pattern matching identifying some useful structures for the recognition of semantic units that are candidates as multi-words. The typical structures searched are NOUN-PREPOSITION-NOUN and NOUN-ADJECTIVE. All that allowed to build a large knowledge base for the automatic identification of multi-words.



NOMI	FREQUE	CODICE	CODIC	CODICE2	REST	FIELD7
accertamenti bancari	4	*	*	ex		accertamento bancario
accertamenti calligrafici	4			nuove		accertamento calligrafico
accertamenti cardiovascolari	1	*		nuove		accertamento cardiovascolare
accertamenti clinici	71	*		nuove		accertamento clinico
accertamenti condotti	22	*		nuove		accertamento condotto
accertamenti contabili	11	*		nuove		accertamento contabile
accertamenti dattiloscopici	3	*		nuove		accertamento dattiloscopico
accertamenti definitivi	30	*		nuove		accertamento definitivo
accertamenti degli investigatori	2	*		nuove		accertamento degli investigatori
accertamenti dei carabinieri	6	*		ex		accertamento dei carabinieri
accertamenti dei finanzieri	4	*	*	ex		accertamento dei finanzieri
accertamenti dei nas	2	*	*	ex		accertamento dei nas
accertamenti dei periti	4	*		ex		accertamento dei periti
accertamenti del caso	10	*		ex		accertamento del caso
accertamenti del ros	2	*	*	ex		accertamento del ros
accertamenti della polizia	4	*		nuove		accertamento della polizia
accertamenti di imposta	18	*				accertamento di imposta
accertamenti di responsabilita'	9	*				accertamento di responsabilita'
accertamenti diagnostici	189	*	*	nuove		accertamento diagnostico
accertamenti disposti	29	*		nuove		accertamento disposto
accertamenti documentali	3	*		nuove		accertamento documentare
accertamenti ecocardiografici	1	*	*	nuove		accertamento ecocardiografico
accertamenti fiscali	50	*	*	nuove		accertamento fiscale
accertamenti genetici	3	*		nuove		accertamento genetico
accertamenti giudiziali	51	*	*	nuove		accertamento giudiziale
accertamenti giurisdizionali	6	*	*	nuove		accertamento giurisdizionale
accertamenti immunologici	2	*	*	nuove		accertamento immunologico
accertamenti impugnati	18	*		nuove		accertamento impugnato
accertamenti incidentali	45	*		nuove		accertamento incidentale
accertamenti incrociati	1	*		nuove		accertamento incrociato
accertamenti induttivi	67	*		nuove		accertamento induttivo
accertamenti investigativi	3	*	*	nuove		accertamento investigativo
accertamenti ispettivi	84	*		nuove		accertamento ispettivo

Figure 3: Multi-word database screenshot.

All the elements identified are collected and associated their frequencies both within the whole corpus formed of the aforesaid texts and within the various sectorial subsets (news, politics, sports etc.), in order to evaluate their degree of dispersion within such subsets. The set of elements identified as valid candidates was inserted in a database and a rank value was associated to each element in order to be able to measure the respective potential strength. An editorial phase allowed the selection – within the database of multi-words – of those compound expressions considered important and pertaining to the field of application. Afterwards, the phase of recognition became a component to be applied periodically for the identification of new elements and, in the case of a positive evaluation, in order to increase the linguistic knowledge base and enrich the resource.

5. Topics

In SP the concept of topic identifies a relevant theme among news concerning a given period. For this reason a functionality was realized that is able to acquire the semantic knowledge of specific sectors/subjects identified by an individual user and to use such a knowledge to automatically select news pertaining to the relative topic.

This one can be selected through the identification of lexical elements that are present in a text (words, entries, named entities, multi-words) or can be identified by means of the manual selection of a subset of articles meeting the requirements expected and then ending up to form the training corpus for that theme.

The approach followed by SP does not make use of a predefined ontology, as it happens in “directory-based” search engines, but recognizes the user’ specific interest by automatically selecting the information expected. This phase takes place off line and can be repeated even more than one time.

The system recognizes the constituent elements of the topic selected among the textual material of the day, week or month (according to specific requests) and provides the user with an ordered list of the articles considered pertaining to the topic expected: substantially, the degree of trust, with which it is asserted that the article deals with that subject-matter, is measured. Various testing phases pointed out a high degree of success of the procedure. However, the greatest problems were discovered in the shortest articles, since the scarce quantity of text did not allowed to correctly weigh the terms (relevant semantically units) in relation to other words of the same text. This one is an intrinsic characteristic of the kind of textual material and requires an careful evaluation of the cases. One of the peculiarities of the system consists in using topics as a further filter in order to personalize the user’s profile. In fact, besides being provided with a series of topics for a general use, the system offers the user the chance of designing his own semantic rule (his own topic), in order to give him back only the information pertaining to the theme selected.

6. Why Is Semantic Press Different from Other On-line Press Reviews?

The automatic selection of elements extracted from texts by means of linguistic and statistical tools in itself offers the SP system a value added. However, the real characterization of the system consists in the capacity of providing individual users with the possibility of adapting the tool to their specific needs. Every user can draw his own profile (even organized into more than one level and one sector) and have a personalized press review. Users can select various typologies of elements in order to sketch their own profiles:

1. sources (one or more headings);
2. sectors of interest (typically newspaper' sections such as news, politics, sports etc.);
3. words, entries, named entities, multi-words etc. (namely all those words that can identify – transversally with respect to sectors – the news that are of major interest);
4. topics.

The last two possibilities, which have an evident linguistic value, are the most important elements to identify interesting information. The system ability to manage profiles is for us its most important characteristic and suggests its application in a lot of sectors of the Public Authority. In fact, different levels of profile can be hypothesized within a public authority. For instance, sectors such as traffic, consumptions and the street cleaning service can be identified by a first level. Subsequently, for each level specific sub-levels can be identified (for instance, for the street cleaning service: the refuse collection, the refuse disposal, incinerators, dangerous refuse etc.). No restriction exists on the creation of new levels and every user of the system operating in the sector can specialize his profile by introducing levels better representing his interests (for instance, by linking them to a particular area of the territory or by adding new sectors connected to more strictly private interests such as politics, sports etc.). In this way, specific subsets of news are extracted from the whole of the articles collected by the system.

The various profiles can generate personalized Web pages as well as the automatic e-mailing, when requested, or the immediate send of SMSs, in case of requests for alerting as far as specific news are concerned.

7. Future Works

The great boost given by the project to the creation of tools and language resources enriching the ones of the LM system should be underlined.

Such abundance of experiences and results is now available for other potential applications and LM is more and more shaping as a powerful tool for Text Mining, classification of textual material, knowledge extraction from texts.

The prospects of development look particularly interesting for the SP project, in which the field of application of the system can be changed: no longer texts with a generic content such as those available in on-line newspapers of public interest, but specific sectors of texts.

The specificity of contexts enormously contributes to a greater productivity of tools and resources. For instance, the applicability of SP to texts of health communication is being weighed to provide all the experts of the sector with a supporting, knowledge and decision tool. Moreover, an area exists, in which SP is already being operatively applied: the one relating to juridical and normative texts. Here texts are characterized by lexical, syntactic and grammatical uniformity and allow the creation of information and alerting systems, which are very important in the eParticipation and eGovernment sectors.

8. References

Effective Information Discovery, Supporting the Analytical Mission through Entity-Based, Semantic Discovery. An Inxight Federal Systems Group White Paper, November 2006.

Installing the Intelligent Miner products: Modeling, Scoring, Visualization V8.2. IBM Manuals for DB2 Intelligent Miner Modeling V8.2.

Pianta, E., Speranza, M., Magnini,, B., Bartalesi Lenzi, V., Sprugnoli, R. (2006). Italian Content Annotation Bank (I-CAB): Person Entities (V.1.3) Tecnical Report, ITC-irst, Trento, Italy.

Federico, M., Bertoldi, N., Sandrini, V. (2002). Bootstrapping Named Entity Recognition for Italian Broadcast News. In Proceedings of the Conference Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, USA, pp. 296-303.

Picchi, E. (1994). Statistical Tools for Corpus Analysis: A Tagger and Lemmatizer for Italian. In Willy Martin, Willem Meijs, Margreet Elsemiek ten Pas, Piet van Sterkenburg & Piek Vossen (Eds.), Proceedings of Euralex '94, Amsterdam, The Netherlands.

Picchi, E., Ceccotti, M. L., Cucurullo, S., Sassi, M., Sassolini, E. (2004). Linguistic Miner: an Italian Linguistic Knowledge System. In Proceedings of LREC 2004, Volume V, ELRA, Paris, France, pp. 1811-1814.

Picchi, E., Cucurullo, S., Sassolini, E., Bertagna, F. (2008). Mining the News with Semantic Press. In Proceedings of LangTech 2008, Roma, Italy, pp.141-144.