Automatic dictionary and rule-based systems for extracting information from text

Sergio Bolasco, Pasquale Pavone
Dip. di Studi Geoeconomici, Linguistici, Statistici, Storici e per l'Analisi Regionale
Università "La Sapienza" di Roma
sergio.bolasco@uniroma1.it, pavone03@libero.it

Abstract: The paper shows a general introduction to the use of meta-information in a text mining perspective. The aim is building a meta-dictionary, as an *available linguistic resource* useful for different applications. The procedure is based on the use of a hybrid system. The suggested algorithm employs, conjointly and in a recursive way, dictionaries and rules, the latter both lexical and textual. An application on a corpus of diaries from the Time Use Survey (TUS) by Istat is shown.

Keywords: automatic classification, meta-data, linguistic resources, hybrid system.

1. Introduction [1]

The importance of meta-data for the extraction of information from texts is undoubted and unanimously agreed upon (Basili, 2005; Poibeau, 2003). Generally, in the field of natural language processing, the meta-data consist of annotations and categorizations of lexical and textual units (Bolasco, 2005). In the present work, a procedure based on a hybrid system is proposed in order to construct linguistic resources that can be re-used - in a perspective of text mining - for the extraction of entities from a corpus of textual data. To this purpose, the integration between the level of both lexical and textual analysis is crucial.

In the lexical automatic processing (LAP), the object of study is the lexicon of a corpus of texts. The unit of analysis of the text is the "word" as a *type*, uniform as it is a lexia (that is, an elementary unit of meaning which is not decomposable further), but also mixed, as it can consist of an inflected form, a polyform or a lemma. The annotations are performed on a "vocabulary" table, which results from the numerical coding of the corpus types (numerilization) and from further re-coding (re-numerilization) of the types caused by lexicalisation, disambiguation, lemmatisation and/or stemming works.

Lexical annotations of *types* (meta-data) can be of various kind: linguistic, semantic, quantitative and statistic. These annotations are produced during different steps of the processing: text normalization, grammatical and/or semantic tagging, index TFIDF calculus. These processing stages will make possible then to extract and select significant vocabulary's parts in order to describe the lexical characteristics of the corpus (for example, what are the significant elements of each part of speech, as verb, noun, pronoun, adjective, adverb, preposition, conjunction and interjection). The selection will even "illustrate" in the factorial maps some sub-sets of lexical units of interest.

Usually, for the purposes of text mining, meaningful parts of the vocabulary are selected: a *peculiar* language (over/under-used with respect to the expected use of a frequence dictionary of reference), a *relevant* language (extracted through the TFIDF

[1] The present research was funded by MIUR 2004 - C26A042095. This paper is the result of the joint work of two authors, Bolasco and Pavone. Sections 1, 2 were written by Bolasco and sections 3 by Pavone.

index (Salton, 1989) which is discriminant among the documents), and a *specific* language (characteristic of some partitions of the corpus). Each of these constitutes an example of meta-information attributable to the lexical units.
In the textual automatic processing (TAP), the object of study is the corpus analyzed as a collection of documents (fragments) to be "categorized". In this case, the automatic annotation on the text can be of one of the following types: *linguistic* (individuation of structures or syntagms with variable elements: Balbi *et al*., 2002), *semantic* (from concepts up to more complex structures such as ontologies: Pazienza, 2003), or *quantitative* (TFIDF with respect to a query), *statistic* (that is the probability of different meanings of the same word). The unit of analysis is the fragment (which can be different from a single phrase of the whole document), as a context-unit.
Textual analysis will be characterized by selecting/extracting significant elements from the investigated corpus. It will be searched from time to time single types or classes/categories of a specific type or even a named entity (name, toponym, company). Relations between types, classes, categories and entities will be searched as well, by establishing searching criteria and rules based on logic operators.
Each information extracted from a text is an **entity** of interest in itself. The entity search is performed by writing a regular expression, which is a typical operative function of text mining [2].
The annotations for the textual analysis are carried out on a "fragments" table that contains both *a priori* variables (the categories of partition of the corpus) and variables that are the result of the TAP. These annotations are the result of a process of Extraction, Transformation and Loading (ETL) capable to search a non-structured information within a text and to transform it into a structured information in a database, useful to be found again during subsequent work phases, rapidly and exclusively. The annotation can be done in several ways: information presence (yes/no), number of time in which it appears, registration of what follows the entity searched within the text. Each query in TAP is performed by a regular expression. Query execution produces a list of the founded entities with relative both frequency within the corpus and distribution in each fragment.

2. A model for creating a meta-dictionary by means of a hybrid system

The meta-data are obtained by models and it is possible to re-use them again through resources. The differences between a model and a resource are defined below.
A model, within the field of automatic text analysis, is a set of "open" instructions which express one or more rules. The model, when applied to different corpus with respect to the ones it is made for, the model produces new but also unexpected results. A lexical query such as ***nipot***, for example, extracts terms such as *nipote/i/ino, pronipote* from a corpus concerning the description of daily activities. In a different corpus, for example in a collection of press articles, the query finds, including the previous ones, even new terms such as *nipotastro, pronipotino, bisnipote, arcipronipote* and it includes *false positives* (noise presence), such as *plenipotenziario, inipotizzabile*. In general, a model gives also the opportunity to retrieve of *false negatives* (reduction of silence), since it recognizes spelling mistakes compatible with the query (*nipotini, nipotiva*).

[2] This function is available in computer programs for the analysis of texts, such as, in particular, TaLTaC2 (www.taltac.it).

A resource, again within the field of automatic textual data analysis, is a set of "closed" instructions defined in a list (dictionary). Each time it is applied, it reproduces itself at the most. Therefore it does not allow for discovering new elements, it does not introduce false positives (absence of noise), but it does not allows either for the discovery of possible false negatives (it cannot reduce the silence)

A **hybrid system** is an algorithm for the extraction of information from a text, characterized by the combined and iterated use of dictionaries (DIZ) and rules (REG). A hybrid system produces as a final result a list of entities (meta-dictionary).
A **dictionary** consists of a list of predefined lexias. When these lexias are polyforms, a new entry in the vocabulary of the corpus is produced upon their recognition (*lexicalization*).
A **rule** defines a condition for an entity search in the text; often, it allows one to identify entities through a correlation between one or more categories and/or types. The application of the same rule to different corpora results in both predictable and unexpected entities: in the latter case, new elements are discovered which are permissible under that rule. However, some entities can be *false positives*, because they are not pertinent with respect to the information being sought. Therefore, they must be eliminated from the list. Examples of lexical rules are queries for the search of lexemes, of infixes and of morphemes in the dictionary of the corpus. Examples of textual rules are queries written by means of regular expressions that combine classes of types obtained from the application of dictionaries via boolean operators (AND, OR, AND NOT).
The application of a dictionary and/or of a lexical rule allows for the annotation with a label of both the types of dictionary and the corresponding tokens in the corpus. The elements that have the same label constitute a *class* and are equivalent to each other, like "synonyms". The meta-dictionary, - come risultato dell'applicazione dell'insieme di parecchi dizionari e regole che definiscono il modello basato su un sistema ibrido - , once controlled and cleaned up to eliminate the false positives, constitutes the resource to be re-applied to textual *corpora* of the same type.
As is well-known, every **model** is created in three stages. A first phase, of construction, is required for empirically determining the basic components of the structure of the model (*training*); that is, the single entries of a dictionary, or the *operanda* of each rule. These are put to test many times on the dictionary and/or the text, until a definitive choice is made. A second phase consists of the formalization of the model, by means of creation of the meta-list and the meta-query (see below). A third phase consists of the application of the theoretical model: it applies the model to the corpus being studied or to other *corpora* of the same type.
The algorithm that organizes *dictionaries* and *rules* (also in a recursive way) into processes that are firstly **explorative** - first lexical (see step (A) below), then textual (B) -, and subsequently, after the model formalization step (C), **applicative** - textual (D) and lexical (E) -, is articulated in the following steps:
A) Predispose classes of types *at the lexical level* by means of uni-label (lists) or multi-label (tables) **dictionaries**, and/or **lexical queries** (uni-label dynamic dictionaries produced from elementary rules on single lexias: prefixes, lexical morphemes, infixes or suffixes).[3] This phase allows one to explore and define the constituent parts of the structure of the model.

[3] Such dictionaries and lexical queries feed with an equal amount of labels the CATSEM field in TaLTaC2.

B) Look for relevant entities through **textual queries** by applying regular expressions *f(x)* that localize sequences of words in the corpus. Each *f(x)* combines two or more of the classes realized at step A, producing a list of individuated sequences, both as entities vocabulary (V) and in terms of positioning of tokens (N).
C) Step of *model formalization as set of rules*. Once the dictionaries, the lexical queries and the single *f(x)*s have been validated, in order to repeat with a single action the annotations in the vocabulary of the corpus; a **meta-list** and a single textual **meta-query** (that collects the single *f(x)*s, so obtaining the model in its total structure) are defined.
D) The **application** of the meta-query makes the **model** up-to-date (*as final list*). This vocabulary of the individuated entities[4] supplies redundant occurrences, because each *f(x)* puts into action an automaton with finite states that scans the text byte by byte and counts all the entities individuated by each single query. Therefore, shorter entities, such as <*house*>, are included in longer ones, such as <*my house*>, <*my mother's house*>, <*relatives' house*> and so on.
E) Re-apply this dictionary of entities, depurated of false positives, and assumed as **meta-dictionary** (*available resource*), for a semantic tagging aimed to lexicalize the entities of the corpus. With such an operation, the occurrences of every entity (as lexias of the corpus vocabulary) are made exact: that is, the tokens of <*house*> do not include those of <*my house*> and the previous ones.

3. Application to the Istat TUS survey

In what follows, an application of the hybrid system is described, which has been carried out on the corpus of 50,000 diaries of the Time Use Survey (henceforth, TUS) - *Istat* 2002-2003. In TUS, each diary is written in free text, and describes the activity performed by a person in the course of the day, according to intervals of 10 minutes (minimum). Contextually, the place and/or means of transport in which the activity takes place are annotated. The corpus amounts to approximately 9 million occurrences (Bolasco *et al.*, 2007). The construction of the model has the objective of characterizing the thousands of locutions used in order to describe the places of the daily activities. These have as their basic linguistic structure a *prepositional syntagm*, composed, in Italian, as follows:

PREPOSITION + (ADJECTIVE) + SUBSTANTIVE + (ADJECTIVE) (1)

The adjectives are placed between parentheses because their presence is optional. For example, as regards the elementary locution "*a casa*" ("at home"), the model recognizes sequences of the type: "*a casa mia*" ("in my house"), "*nella mia seconda casa*" ("in my second house"). The structure can be found, even several times (the examples below show each one separated by the symbol |), with adjectival function with respect to the main substantive (e.g., "on the seat | of the machine"; "to the party | of birthday | of a friend"). In the diaries contractions such as "*vicino casa*" (standing for "vicino a casa") can also be found. Table 1 illustrates the typology of space locutions relative to the entity "means of transport".
In the exploratory stage, the basic constituents of the model have been defined, preparing the dictionaries (see table 2) composed of: the list of prepositions; a multi-label table of adjectives, distinguishing between possessive and qualificative; and lexical queries regarding the substantives.

[4] This list of the entities contains a lot of 'compatible trash' ("me house", "o the table", "inthe bed", "in a rooms"), and consequently 'grasps' the phenomenon fully, beyond spelling and grammar.

Table 1: *Some examples of the structure of the model of prepositional syntagms.*

PREP1	ADJ		SUBST	PREP2	ADJ	SUBST
in			*auto / macchina / treno automobile / autobus*			
dentro / fuori / presso la, davanti / vicino alla			*macchina*			
nella	*mia*		*automobile*			
sul		*nuovo*	*autobus*			
in			*macchina*		*mia / sua / loro*	
sull'			*auto*	*di un*		*vicino*
verso la			*fermata*	*dell'*		*autobus*

Table 2: *Some elements of dictionaries*

PREP1	POSS	ADJ	SUBST	PREP2	POSS	ADJ	SUBST
da un	*loro*	*altro*	*auto*	*degli*	*mia*		*amica*
dal	*mia*	*nuovo*	*autobus*	*dei*	*mio*		*amici*
dall'	*sua*	*nuova*	*autocarro*	*del*	*sua*		*amico*
dentro il			*automobile*	*dell'*	*suo*		*amiche*
dentro l'			*autostrada*	*della*			*azienda*
in			*autovettura*	*delle*			*collega*
in un			*macchina*	*di*			*ditta*
nel			*moto*	*di un*			*figlia*
nell'			*motocicletta*				*figlio*
nella			*motopeschereccio*				*mamma*
su un			*motorino*				*nipote*
su un'			*tram*				*nonni*
sul			*treno*				*nonno*

The construction of these elements has been performed according to various criteria: the prepositions have been categorized in different ways on the basis of their position in the structure[5]; and substantives and adjectives have been individuated by applying lexical queries based on a reduction to lexical or grammatical morphemes (as an example: auto*, moto*, *ary, *teque). With these queries, unpredictable entities have been obtained, by adding both elements compatible with the rule (as an example, from auto*: *autobus*, *autocar*), and false positives (*autobiographic*, *autogestion*).
The model has been completed by means of the repeated application of textual queries written with regular expressions aimed to reconstruct specific parts of the structure of the graph. For instance, using sub-lists of prepositions for some substantives, locutions of place are only individuated when supported by those prepositions (e.g.: <*to the pub*>, <*to the doctor*>, so inhibiting sequences, such as <*of the pub*> or <*by the doctor*>, that are not locutions of place).
The graph in figure 1 formalizes the definitive model expressed in (1).

In the second stage, on the basis of this graph, a single meta-list (table 3) and a single textual meta-query are reconstructed. In more detail, such a query was composed by a regular expression consisting of 39 sequences in "OR"

(e.g.: "PREP1 SUBST" OR "PREP1 ADJ SUBST" OR "PREP1 POSS ADJ SUBST" OR ... OR "PREP1 ADJ SUBST PREP2 ADJ SUBST" ...).

[5] The list PREP1 contains the main (simple, articulated and improper) prepositions compatible with the sense of the prepositional structure; the list PREP2, instead, contains only the simple and articulated form of the preposition "of".

Figure 1: *The formalization of the model*

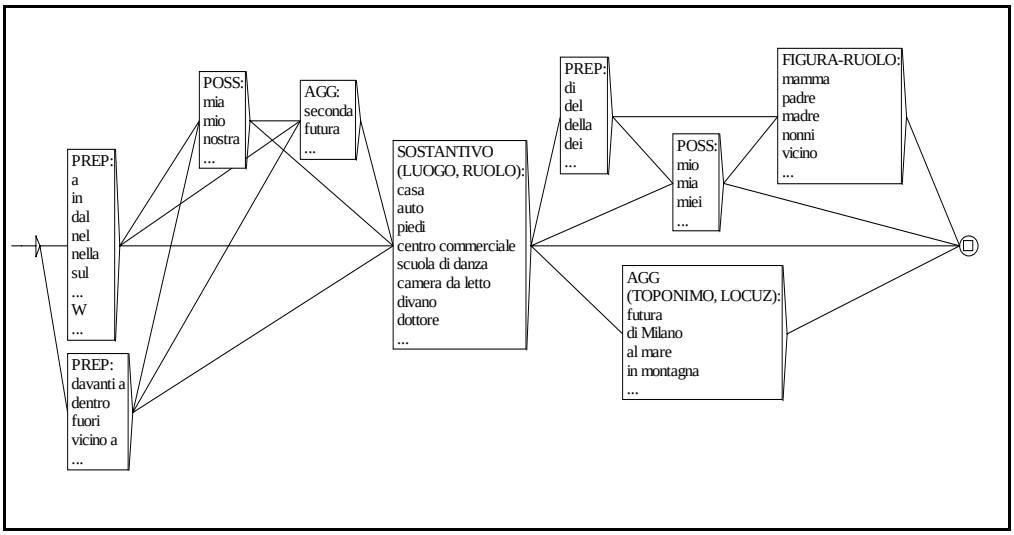


Table 3: *Example of the meta-list*

type	label	type	label
a	prep1	*casa*	subst
alla	prep1	*auto*	subst
da	prep1	*...*	
dal	prep1	*di*	prep2
...		*del*	prep2
mia	poss	*...*	
sua	poss	*futura*	adj
...		*nuova*	adj

Only then has one moved onto the third phase, that is, to the application of the meta-query. This application has individuated 6,388 entities, for a total of 1,731,630 'gross' (redundant[6]) occurrences. The entities have been cleaned up to get rid of the false positives, obtaining as a final result 5,404 locutions of place. These will constitute a reference point (meta-dictionary) for any future survey. By applying this *resource* for a semantic tagging, the 5,404 entities of the TUS corpus are lexicalized (table 4).

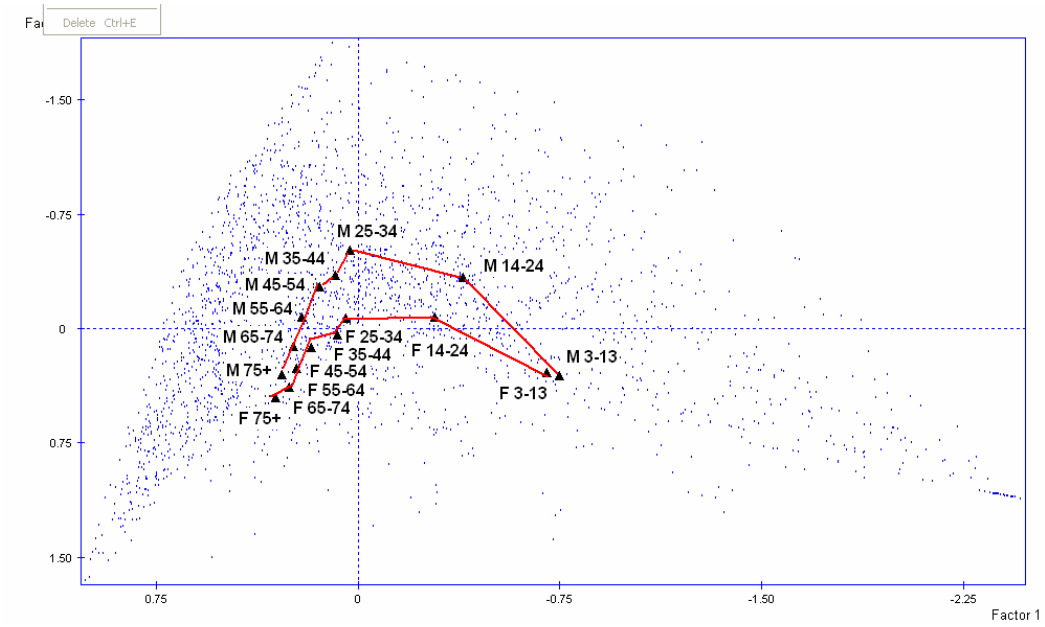
Table 4: *Some examples of locutions of place*

Locution	Occurrences	Locution	Occurrences	Locution	Occurrences
a casa mia	*377.866*	*sul divano*	*7.344*	*nella mia cameretta*	*90*
a piedi	*72.428*	*in ufficio*	*5.481*	*su una panchina*	*88*
in macchina	*43.712*	*in spiaggia*	*3.347*	*nel cortile della scuola*	*64*
a letto	*38.113*	*in giro*	*2.161*	*ad una festa di compleanno*	*48*
in cucina	*18.766*	*nell' orto*	*2.145*	*in mezzo alla natura*	*35*
al bar	*15.169*	*presso la propria abitazione*	*320*	*vicino al caminetto*	*32*
a scuola	*14.880*	*alla fermata dell'autobus*	*290*	*sulla sedia a rotelle*	*24*
in bagno	*14.684*	*dal giornalaio*	*233*	*fuori dal mio paese*	*15*
al lavoro	*11.244*	*davanti alla tv*	*202*	*verso il centro commmerciale*	*11*
per strada	*10.094*	*sotto l' ombrellone*	*186*	*tra i negozi dell' ipermercato*	*2*

[6] see step D in the paragraph 2.

In general, the results of the above-mentioned queries, as pointed out in the paragraph 1, produce new variables that are inserted in the matrix of the fragments (individual diaries). These variables constitute a representation of the "concepts" or relations among concepts that are to be correlated to the a priori information (e.g., the structural variables of the individuals).
In this case, it is possible to emphasise the correlations between the locutions and individual characteristics via factorial analysis. The latter allows one to reconstruct in detail the relationship between the various kinds of locutions and the individuals, by partitioning the corpus of the diaries according to age × sex. From the overall analysis of all the locutions (a matrix $n \times p$, where n=5404 locutions and p=16 classes age × sex), such strong relationships emerge that the resulting map – shown in figure 2, where each point individuates a locution and the barycentres of the age × sex classes are connected by a line – can be described according to the slogan "Each age has its places".

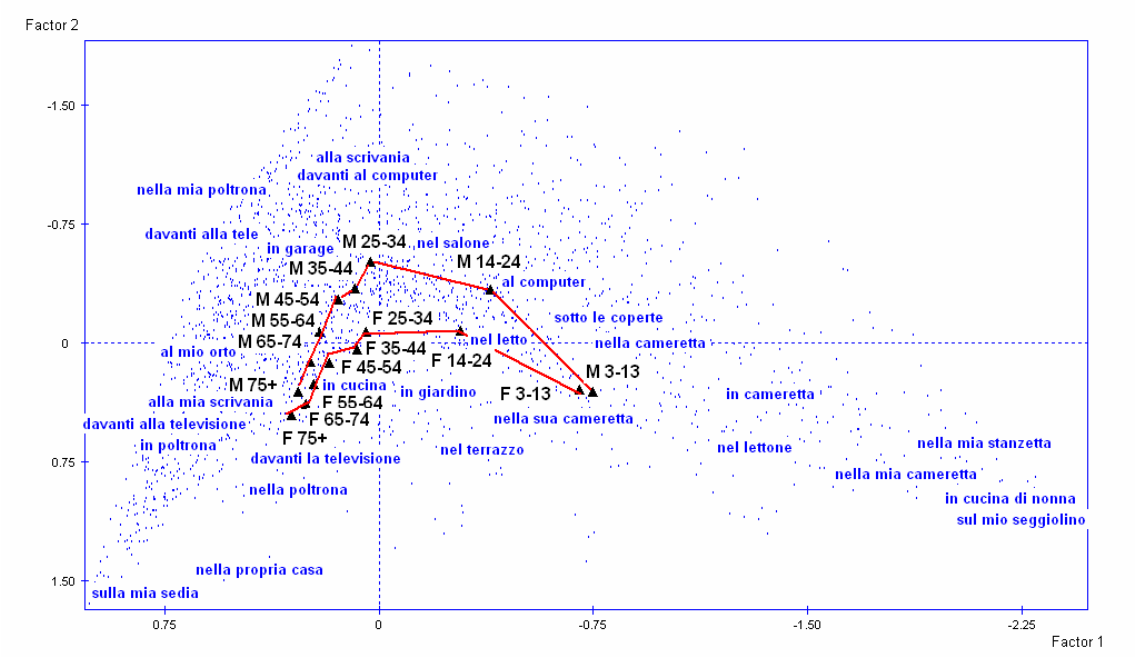
Figure 2: *Factorial analysis of the locutions of place by age×sex groups - TUS 2002-2003*



As can be observed in the factorial plane, in the young age there is a marked variability of places; the latter increases as age increases (the maximum is reached around the age of 20-25, in proximity to the origin of the factors), and then decreases as the old age approaches.
If we go into more detail, let us consider the thematic list with reference to "places inside one's house" (figure 3). The maximum variety of places during the day occurs for the age groups 'in-between' (*alla scrivania, davanti al computer, nel salone, in giardino, ..., nel terrazzo*), while as the years go by mobility (which begins in the early years: *sul mio seggiolino, nella mia cameretta*) gets more and more limited (*davanti alla televisione, in poltrona, davanti al camino*) and eventually disappears (*sulla mia sedia*).
It is interesting to note how the differences between the sexes gradually increase towards middle age (see figure 3, *M: in garage; F: in cucina*), then tend to disappear in older ages (*nella propria casa*). Furthermore, the barycentre (of each class of the women) is slightly more to the left, that is towards the older ages, which is consistent with the greater life expectancy of the female sex.

The 5000 expressions of place, although so many can not represent the very all places. If we consider the place "sea", the TUS corpus provides: at sea, over the sea, on the seafront, close to the sea, on the seashore, in a village of sea, the home of the sea, etc.. But not a frequent place as "at the bottom of the sea". Therefore they are just one exhaustive example of places of everyday activities. As such this meta-dictionary is a re-usable resource for the next Time Use Survey 2007-2008 of Istat.

Figure 3: *Factorial analysis of the locutions of place with reference to "places inside one's house" according to age × sex groups – TUS 2002-2003*



References

Balbi S., Bolasco S., Verde R. (2002) Text Mining on Elementary Forms in Complex Lexical Structures in A. Morin, P. Sébillot (eds.) *JADT 2002*, St. Malo 13–15 marzo, IRISA-INRIA, 89–100.

Basili R. and Moschitti A. (2005) *Automatic Text Categorization. From Information Retrieval to Support Vector Learning*, Aracne, Roma.

Bolasco S. (2005). Statistica testuale e text mining: alcuni paradigmi applicativi, *Quaderni di Statistica*, 7, 17–53.

Bolasco S., D'Avino E., Pavone P. (2007) *Analisi dei diari giornalieri con strumenti di statistica testuale e text mining*, in M. C. Romano (ed.) "I tempi della vita quotidiana. Un approccio multidisciplinare all'analisi dell'uso del tempo" Roma, ISTAT, 309–340.

Pazienza M.T. (Ed.) (2003) *Information Extraction in the Web Era. Natural Language Communication for Knowledge Acquisition and Intelligent Information Agents*, LNAI 2700 Tutorial, SpringerVerlag, Berlin.

Poibeau T. (2003) *Extraction automatique d'information*, Hermes–Lavoisier, Paris.

Salton G. (1989) *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*, Addison–Wesley.