An Operational Approach for Document Content Characterization in Specialized Domain.

Antonino Mazzeo
mazzeo@unina.it

Prof. Antonino Mazzeo is a Full Professor at the University of Naples 'Federico II', Italy. He has led research projects partially supported by the Ministry, CNR, ASI and EC. He is involved in scientific collaborations with international research agencies and universities. He has a wide experience in the field of information systems, applied to the PA domain.

Flora Amato
flora.amato@unina.it

Flora Amato is a PhD student of Computer and Control Engineering at the Department of Computer and Systems Science of the University of Naples 'Federico II', Italy. Her current research interests lie in information retrieval, natural language processing, knowledge extraction and management.

Rosanna Canonico
rosanna.canonico@unina.it

Rosanna Canonico works at the Department of Computer and Systems Science, University Federico II of Naples, where she collaborates to many projects about semantic processing of documents, semi-automatic analysis of text, Information Retrieval and Information Extraction.

Abstract

Several problems prevent documents in natural language to be comprehended through automatic procedures, first of all, morpho-syntactic, syntactic and semantic ambiguity. Generally, ambiguity derives from the dynamism and the flexibility of language signs which can acquire new uses and, consequently, add new senses to their meanings in order to meet specific communicative requirements. We need, in fact, to consider the complexity and the inter-dependence of the syntactic, semantic and pragmatic processes involved in the comprehension of documents in natural language. A document is the product of a communicative act where different factors come at stake, it is a semiotic object resulting from a process of collaboration between its author and a reader: the former encodes by means of language signs the intended meanings, the latter decodes these signs resorting to the knowledge of the language and of the infra-textual context and interprets them by resorting to the knowledge of the extra-textual context and, more in general, to his encyclopedic knowledge. Reducing the analysis to documents pertaining to specialized domains (examples of specialized domains are Public Administration, Judicial System and, more in general, the bureaucratic field), we can reasonably state that the interpretation of the signs is unique, since the limits introduced by the domain lead to a formal definition of the interpretations related to the signs; which means that, in many cases, the process of coding/decoding can be finalized without ambiguity. In this paper, therefore, we first propose a semiotic model for document characterization, in order to describe general properties and characteristics of documents that could be customized to specific cases; moreover, we propose a semi-automatic methodology for document content characterization for the automatic comprehension of parts of document in specialized domain. We do not claim to characterize the content of the whole document: our aim is to individuate the peculiar concepts of the specific domain of reference.

1. Introduction

The increasing amount of paper and digital document sources, both in the public and the private sector, has nowadays pointed out the need of a deep renewal of the old information systems for a better and a more effective access to the document contents: the automation of document management procedures is more and more required and, in this field, techniques of semantic processing seems to be the only ones giving a surplus value to all the procedures requiring access to multimedia document collections.

Every day various categories of users need to access to huge quantities of contents embedded in documents of different formats and typologies: the capacity to retrieve only the relevant documents and to extract data from them, the possibility to transform these data into relevant information and to represent them in an opportune way (according to the necessities of the users and the current laws), are key parameters for an intelligent knowledge management, above all in terms of information accessibility and sharing, as well as long-term preservation.

Firstly, an intensive and extensive activity of *dematerialization* is required: the digitalization of paper documents and archives, with the consequent automation of the procedures of document management, enables a reorganization and an improvement of the efficiency and effectiveness of the information systems. Secondly, strategies of *text analysis* and *extraction of relevant information* are required in order to provide a terminological and conceptual representation of documents, aimed at a semantic retrieval.

The automatic derivation of knowledge from texts is a fundamental task in the Semantic Web scenario but it is also difficult one, due to the fact that knowledge is rarely explicitly expressed. The state of the art in this field involves techniques of Natural Language Processing (NLP) and a cross-disciplinary perspective including Statistical Linguistics (De Mauro, 1961; Rizzi, 1992) and Computational Linguistics (Chiari, 2007; De Mauro, 1994; Gallino, 1991; Gigliozzi, 1997; Lenci *et al*., 2005; Orlandi, 1990), whose objective is the study and the analysis of natural language and its functioning through computational instruments and models. In particular, for the analysis of limited textual universes, such as specialistic areas, specific disciplines have developed, like Corpora Linguistics (Spina, 2001) and Textual and Lexical Statistics (Bolasco, 1999; Bolasco, 2004; De Mauro, 1997; La Torre, 2005).

Texts in natural language can be therefore studied by means of hybrid systems combining the traditional techniques of language analysis with statistical techniques: if from one side Linguistics enables the description and the analysis of the document linguistic structure, from the other side Statistics enables the identification of the relevant language phenomena, not immediately detectable[1], which serve for the extraction of the peculiar information with respect to the contents dealt in the document and, more in general, to the domain of reference.

The automatic derivation of knowledge from texts is, therefore, nothing but an analysis of the linguistic structure of the text, in order to recreate the ontological model the author of the document has tried to conceptualize and to make explicit within the document itself by means of a system of designations. Obviously, setting an opportune correspondence between the linguistic representation of the ontological domain, as it is realized within the document, and the domain itself, implies the availability of a considerable quantity of knowledge with respect to the field and the themes dealt: the extraction of information from a text requires other information. Knowledge is organized and acquired by means of abstractions named concepts. A concept is, indeed, an abstract unit of knowledge whose definition should ideally include:

- an *intensional meaning*, defined by the set of intrinsic properties that are necessary and sufficient to characterize the concept and make it different from other concepts;
- an *extensional meaning*, defined by the class of referential entities the intrinsic properties of the concept apply to;

Consequently, the comprehension of a particular concept within a specialized domain requires information about the properties characterizing it, as well as the capacity to identify the set of entities the concept can refer to. The more articulate is the intensional meaning of a concept, the more reduced is its extension.

A system of NLP for automatic derivation of knowledge from text confronts therefore with an *interpretative objective*: text data are analyzed to be comprehended and transformed into meaningful information, that is to say into reusable information for retrieval purposes.

However, the automatic comprehension of text data involves a series of problems, first of all morpho-syntactic, syntactic and semantic ambiguity, generally deriving from the dynamism and the flexibility of the language signs which can acquire new uses and, consequently, add new senses to their meanings in order to meet specific communicative requirements. The meaning of text data, in fact, comes from a complex and strong inter-dependence among syntactic, semantic and pragmatic aspects: in order to describe a document and understand its contents it is necessary to identify not only the single signs but even the relations these signs keep up among them, firstly at a syntactic and

[1] An example is given by the statistics that can be produced on nouns, verbs or in general on the significant classes of words in order to point out the most frequent items or obtain a description of the text in terms of incidence of word categories.

semantic level and, secondly, at a pragmatic level, that is to say the relations the signs have with the external context and in general with the proper domain the document pertains to.

A document is the product of a communicative act, a semiotic object produced by a human intervention and resulting from a process of collaboration between an author and a reader: the former makes use of language signs to code the intended meanings, which compose the contents of the document itself; the latter decodes these signs resorting to the knowledge of the language and to the knowledge of the infra-textual context and interprets their semantic contents by resorting to the knowledge of the extra-textual context and, more in general, to his *encyclopedic knowledge*. The text sense is, in fact, determined by a series of concepts which are linked to the general experience or cultural knowledge, learned even by other texts, and that are conveyed by means of language expressions related together.

In specialized domains, the extensional meaning of concepts is more restricted since more specialized and technicalized is their intensional meaning. Here, the interpretation of the document contents is always dependent from the competences and knowledge shared between the author of the document and the reader, therefore we can reasonably state that the interpretation of the signs is unique: the limits introduced by the domain lead to a formal definition of the interpretations related to the signs; which means that, in many cases, the process of coding/decoding can be reached without ambiguity.

In this paper we first propose a semiotic model for a general characterization of documents, in order to describe intrinsic and extrinsic document properties that could then be customized to specific cases: we start from the general notions of text and document, with particular focus on the active role of the receiver as text interpreter; we then propose a semi-automatic methodology for automatic derivation of knowledge from texts in natural language and pertaining to specialized domain with the objective to comprehend parts of the documents themselves.

We do not claim to characterize the content of the whole document: our aim is individuate the peculiar concepts of the specific domain conveyed within the document. These procedures, carried out with meticulousness and accuracy, will permit to obtain, from a confuse set of data, a series of well-structured information.

2. A semiotic model for a general document characterization

A text is an abstract structure, a tissue of heterogeneous elements that in their whole, and on the base of the relations they set among them, represent some contents. More formally, a text can be defined as the *output of a communicative act*, expression of the use of the language in concrete situations: a text is then a realization of the language put into action. It represents a complete and well arranged expressive total unit, produced in a certain way by a sender, who gives guarantee of its semiotic construction, and interpreted in a certain manner by a receiver, who gives guarantee of its semiotic action and is disposed to use the text as information source, having as background a specific social and cultural context.

The text total unit is then given by parts put cohesively and coherently together: the single items are connected because of their meaning, their formal aspects, for the communicative intentions of the author and for the contents caught by the reader, and finally because of the collaboration put into action by these two individuals. Thus a text acquire sense through the intervention of two factors: i) the *infra-textual context*, which makes the text a cohesive structure made of different components linked together by morphological, syntactic and semantic relations; ii) the *extra-textual context*, which recalls the communicative situation, the competences shared between author and reader, the

encyclopedia made of knowledge, traditions, visions of the world, the social and cultural scenery where the text is produced and used, the conceptual domain of pertinence.

A document originates when a text merges with a material support: the concrete representation of the text contents is, in fact, made possible by what is called *res* from which originates a material object that is defined document (La Torre, 2005).

The word "document" identifies any object able to demonstrate (or prove) something by providing people with some kind of information.

The notion of document was for the first time introduced by Carnelutti (1975:86) who defined the document as something corporal providing knowledge about something else: a document is " *cosa che docet […] cioè che ha in sé la virtù del far conoscere".* Carnelutti's definition focuses on the material aspect of the document: it is a *res,* that is a real support that lets people know a fact. Any material can work as support for a document: paper, wax, clay, stone, metal, magnetic tape, film, x-ray, photo negative, optical disk, etc.

However, to become document, a *res* must be *signata*: the material object can represent something else only by means of graphic signs impressed on it. The quality of document is then connected to its *representative capacity*: this capacity is tightly linked to the individual will to attribute a certain meaning to the *res*. Documents, in fact, don't exist in nature: some objects are documents from the very first time of their formation because they are created with the intention of representing certain facts (this is the case, for instance, of official documents); other objects, instead, can become documents in other times, for example when required for a legal use to prove something.

Signs serve, therefore, to make document contents available in order to realize the interest of the author to let other people know and the interest of the reader to know.

In the last decades, digital and electronic documents have more and more substituted analogic documents without losing their informative structures: as a matter of fact, they have, as paper documents, all the appropriate characteristics for receiving and preserving those signs which represent a certain external reality. In particular, the expansion of automatic procedures in document management within public and private administrations has had as effect a progressive process of substitution of the traditional paper supports with digital supports, with the consequent rise of the electronic and digital documents, to which the current laws have confirmed full legal value[2]. This process has certainly caused a deep upheaval, above all in the terminology to use: the expressions "digital document" and "electronic document" are, in fact, often used as synonyms without knowing the real technological meaning of these objects[3].

Current developments in technologies have, in addition, expanded the representative capacity of documents giving the possibility to merge in a same document, and in a more effective way, texts in different formats, where information is then conveyed by elements belonging to different media components[4], which can be written, oral, iconic, video and work in more or less explicit ways[5]. This

[2] To see for example: art. 15, comma 2, L. 15 marzo 1997, n. 59 (the so-called Legge Bassanini-uno); D.Leg. 7 marzo 2005, n. 82 (Codice dell'Amministrazione Digitale).

[3] The expression "electronic document" is to be used to identify originally paper documents which are then electronically acquired: an example is given by paper documents acquired by means of techniques of *Optical Character Recognition* (OCR) and automatically archived fulfilling the laws relating to the substitutive archiviation. On the contrary, the expression "digital document" is to be used to identify those documents which are from the very first moment created by means of digital tools: this is the case, for example, of a document electronically typed and stored on optical disks.

[4] In this case, the medium is not only the tool used to impress on the *res* the signs conveying the meanings but it is also to be regarded as the language used for the representation of the meanings themselves.

[5] Each language has its specificities. Verbal language shows a representation of the facts only after a process of mental elaboration: it serves for argumentation and for a hierarchical organization of the contents, since it segments the contents into words, phrase structures, sentences, building coherent relations among these components. The iconic and visual language, instead, has a pre-analytic quality since it enables the receiver to catch in a simultaneous way what it represents

use of different media has turned the document into a powerful communicative and heuristic object: a semiotic object characterized by different, but coherent, dimensions of content representation.

Furthermore, changes are to be seen even in the way to arrange the exposition of the contents: the traditional linear and sequential arrangement has given way to an associative and hyper-textual one. The capacity of the sign as continuous reference to other signs has been greatly strengthened by electronic tools, able to (virtually) cancel the distance between data and to propose non-linear and flexible strategies for the use of the contents themselves[6].

This new typology of document represents a model where knowledge is regarded as a net of interconnected cultural units, where everything is open to new interpretations. Because the elements belonging to this system are expressed by means of different media, the term hypermedia seems to be more appropriate to describe this innovative way to organize contents[7].

The fundamental element in these systems is the so-called "link" that is nothing but a sign able to direct the reading to other signs which are in the same document or even in other documents. A link is, therefore, a sign recalling other signs: the contents of a hypertext are then defined by the inter-relations of these signs and the role of each element is set on the base of these connections whereas these connections together contribute to determine the global sense of the document.

However, a hypertext makes safe the role of the author: he has to arrange his text, enable certain passages and forbid other ones; the reader, instead, has to activate those connections he feels like more relevant for his needs, but the pertinence of a connection is always set by the author during the hypertext planning. The reader can intervene within the structure created by the author, by deciding the order of his reading and proceeding according to his informative needs: the sense of the text is then built according to the intentions of the author but through the active choices of the reader.

2.1 Author and reader: the role of the interpreter

A text activates a process of cooperation between an author, who constructs the meanings of the text by means of signs, and a reader, who puts into action these meanings by acting as interpreter.

A text, in fact, is incomplete without the intervention of a reader who fills its blank spaces with his grammatical and encyclopedic knowledge, as well as his inferential activity: a great amount of information is let implicit and a reader has to extract them from the text on the base of his knowledge.

Reading, then, puts into action the communicative function of a text, i.e. its capacity to indicate meanings, and through reading text meanings come out from their potential state to become meanings in act: a text exists aside from reading but only through reading a text begins to mean and to communicate. Reading becomes, therefore, a hermeneutical activity, an interpretative act: the reader finds himself between the need to understand the meanings coming from the signs and the need to compare these meanings with his system of knowledge. Furthermore, reading is an abductive activity: a reader is called to make hypothesis of sense about the signs and to subject these hypothesis to a process of verification or confutation. While reading, a reader gives sense to the expressions he

---

and communicates, both on the levels of concepts and sensations: this kind of language is informatively more effective and immediate, for its capacity to attract the observer's eye and to connect the information with images which already exist in his visual memory. Visual elements, therefore, have more relevance for the immediacy of their communicative capacity. Further advantages are given by audio-video which, merging the clarity and the strength of the verbal language with the strength of the image and of the sound, turns the document into an efficient communicative object.

[6] Hypertextual writing must be considered not only in its technical nature but even in its mental nature. A deep link, in fact, binds thought and writing: writing is linked to mental capacities whereas the tool used to write inevitably influences human cognitive functions and the ways to acquire and organize knowledge. But writing (and reading as well) has always a mental dimension because it involves a series of connections with other elements, which are relevant for the construction of the sense. We cannot ignore that our thoughts are made of concepts which are associated in a complex net.

[7] In the common practice, the two terms hypertext and hypermedia are often used as synonyms.

encounters, which compress in more and more complex items, before arriving at the complete realization of the global text content.

The text sense is then determined by the implicit and explicit discourse structures and by the reader's strategies of interpretation, based on his linguistic and semantic-encyclopedic competence. Eco (1979) states that when producing a text, an author must keep in mind the competences of a potential reader. This reader, according to Eco, is not a real reader but some abstraction, an ideal reader, the model of a text strategy: this reader represents the set of the necessary competences to succeed in the comprehension and in the interpretation of the text itself. An author, therefore, determines the form of his text by considering not only the contents to convey but even the point of view of the reader, making inferences on his possible beliefs and expectations, which permits a pre-decoding.

The reception of a text, according to Eco, is tightly linked to the encyclopedic knowledge of the reader, that is a model of knowledge considered as a net of interconnected cultural units: as a matter of fact, meaning is determined by a potentially unlimited set of concepts which are connected to our general experience of the world and to culturally pre-definite structures apprehended through the time from other texts. Each expression can be subjected to a particular interpretation and each interpretation opens to new expressions which are from their side subjected to interpretation: the encyclopedia should then provide with instructions in order to productively interpret a sign in all its possible contexts of use; it should contain the set of all the possible interpretations. The encyclopedia identifies, then, with a semiotic postulate: it is a global competence but it is also structured in levels since different users have it in a different way. Each individual has, in fact, portions of encyclopedia which activate in certain circumstances, on the base of specific contextual needs, in order to interpret certain texts: an interpreter doesn't need to know the whole encyclopedia but only parts of it, those parts which are required for the comprehension of the text itself. The interpreter must then chose the portion of encyclopedia to activate to afford that particular text: in this sense, the encyclopedia turns into a regulative hypothesis which serves to arrive at the interpretation.

2.2 The structuring of the document contents

A document presents as a complex structure containing different types of information organized in different and inter-connected levels: the level of the text *macro-structure*, coinciding with the text surface and defined by the so-called para-text, graphic and stylistic elements[8], and the deeper level of the text *micro-structure*, defined by the single signs and their combinations and relations in the syntagmatic and paradigmatic dimensions.

A text is then the result of different components forming a whole because of their relations: the deep structure, more analytic and abstract, constitutes the semantic level of the document and gives grounds for the coherence and cohesion of the surface structure, more synthetic and global; that is to say that surface elements find full significance only in the discourse where they are inserted.

Macro-textual elements act at an unconscious level and lead the processes of interpretation through a logic of agglomeration and synthesis which plays on the visual faculties of the receiver: for example, a font with different color and dimension or the use of an image or even a particular location of the various elements contribute to direct the receiver's attention on specific contents and, highlighting the internal structural divisions, simplify the retrieval of certain document parts. All these elements give a pragmatic dimension to the text ensuring a better interpretation. The para-text, in particular, represents

[8] For example: title page, title and paragraphs, which serve to give structure to the text; images, drawings and photos, which serve to show, evoke, motivate; graphs, to represent quantitative data; diagrams, to represent structured relations; tables, to visually distinguish elements; fonts, colors, dimensions, which serve to give visibility and importance to specific contents.

the area of first contact for the receiver: he receives from para-text elements indications about the kind of document and the portion of encyclopedia to activate in order to interpret the document itself.

The text micro-structure identifies, instead, the level of the single units of analysis: graphems, words, phrase structures and sentences. The set of these elements defines on the base of the reciprocal relations on the syntagmatic and paradigmatic axis, involving the morpho-syntactic, syntactic and semantic dimensions .

To conclude, document contents are coded both in the linguistic structure and, implicitly, in the global structure by typographical and stylistic conventions. These elements play an important pragmatic role since they permit to cling the single text data to a specific interpretative context, becoming information source.

3. Problems due to natural language ambiguity

The main objective of document semantic interpretation is an effective information extraction so to ensure an efficient retrieval of the documents themselves. This interpretative objective can be realized through an analysis affecting, in particular, the micro-textual level of the document.

However, several problems prevent documents in natural language to be comprehended through automatic procedures, in particular problems due to the ambiguity and the indefiniteness which make expressions compatible with various interpretations. Ambiguity can affect all the levels of the language, in particular the morpho-syntactic, syntactic and semantic ones.

At a first level, there could be problems affecting part-of-speech tagging: given a sequence of words, each word can be tagged with different categories (Tamburini, 2000).

Il	**successo**	**fa**	**male**
Article	Verb: PastPart	Verb	Adjective
	Common Noun	Noun	Noun
			Adverb

Table 1: Example of word category ambiguity for the Italian language

In the example above, the disambiguation of a lexical item is enabled by the linguistic context (for example, the word "successo" is disambiguated as common noun since preceded by an article), by taking into account the POS category of the preceding or following words. However, it is also to take into account that even the preceding word can be ambiguous or that the disambiguation of a form can require further semantic or pragmatic knowledge.

Automatic POS tagging is a general problem of *word-category disambiguation* involving two kinds of difficulties: i) finding the POS tag or all the possible tags for each lexical item; ii) choosing, among all the possible tags, the correct one. The first problem can be solved by using a glossary or a lexical list as reference; the second one, instead, can be solved by using: i) contextual evidences, that is examining the context where the word is used (linguistic approach); ii) probabilistic evidences starting from a tagged corpus to be used to train a tagger (statistical approach)[9].

[9] Many researches have been conducted on the problem of automatic pos tagging and different have been the approaches used (linguistic, statistical and hybrid) and the models implemented. Among the principal techniques are: *stochastic models* (Charniak *et al.* 1993; Carlberger, Kann 1999, Dermatas, Kokkinakis 1995, Derose 1988; Kupiec 1992), *rule-based models* (Voutilainen 1995), *hybrid systems* (Brill 1995), *memory-based models* (Daelemans, Zavrel 1996), *decision trees* (Màrquez, Rodrìguez 1997; Schmid 1994). Brill e Wu (1998) combine the output of different taggers to obtain the best performance by means of a vote mechanism: for each word is selected the tag that has been chosen by the higher number of taggers (*majority voting*). Among the works developed specifically for the Italian language are De Mauro *et al.* (1993), for stochastic taggers, and Delmonte *et al.* (1997) for rule-based taggers.

At a syntactic level, there are problems affecting the disambiguation of syntactic structures: it is to note that some sentences are, in fact, susceptible of different interpretations, that's why they can be associated to different *parse trees*. This is the case, for instance, of an Italian sentence like "La vecchia porta la sbarra" to which two parse trees can be associated (Figure. 1) since interpretable in two ways: i) an old woman brings a bar (1st parse tree); ii) an old door blocks something (2nd parse tree).

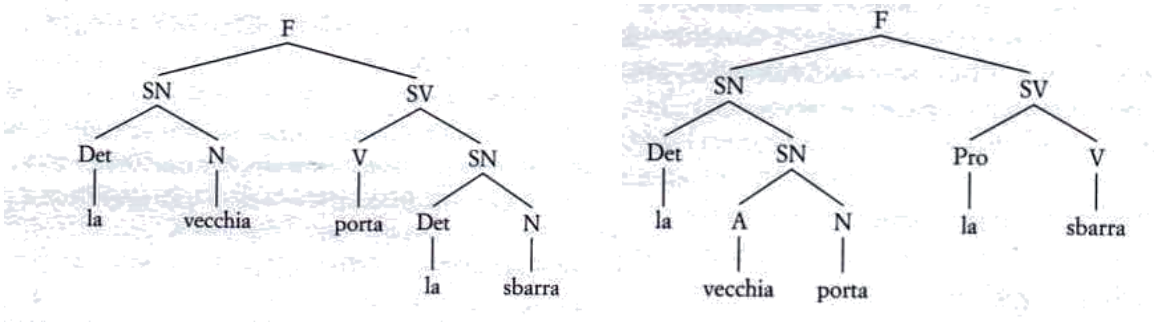

Figure 1: Example of syntactic ambiguity. F (ITA: frase) corresponds to the English S – sentence -; SN (ITA: sintagma nominale) corresponds to the English NP – noun phrase -; SV (ITA: sintagma verbale) corresponds to the English VP – verb phrase -; Det, N and Pro stand respectively for determinative, noun and pronoun.

At a semantic level we have firstly to consider the *unpredictability* with which the word meanings develop and get organized. Meanings are internally organized in senses and very often the senses of a same word get specialized in very different and unpredictable ways, Another aspect related to the organization of meanings is their *extensibility*, that is the capacity to develop for a same word new senses to its meaning in order to meet specific communicative requirements.

Secondly, the presence of homonyms or polysemous words is another aspect representing a problem for interpretation in a computational field. Homonyms are words which are characterized by a common signifier relating to different senses, which are neither etymologically linked nor derivable the one from the other [10]. A word, instead, is polysemous if different senses are associated to it: these senses are all etymologically linked and derivable the one from the other[11]. If for a human interpreter these characteristics are normal and easily to manage, for a computer the matter is different since the management of these issues require a great quantity of elaboration to implement operation of disambiguation[12].

[10] It is noteworthy to differ *absolute homonyms* from *textual homonyms.* The first ones are words having the same spelling (homographs) and the same phonetic form (homophonics). They belong to the same part of speech and often to the same inflectional class: some Italian examples are "calcio" considered as "kick" and "calcio" considered as "pistol grip" (both of them are male nouns with singular form in –o and plural form in –i). The second ones, instead, are words belonging to different parts of speech and to different inflectional classes but similar in the spelling and in the phonetic form: some Italian examples are "faccia" considered as "face" and "faccia" considered as form of the verb "to do". In this case, the words are homonyms only in some of their possible textual forms and not in all of them (Chiari 2007).

[11] Consider, for instance, the Italian word "parte" having different senses, such as: "sezione o porzione" (Engl. "section or share"), "direzione" (Engl. "way"), "parte in causa" (Engl. "party"), "ruolo in una rappresentazione scenica" (Engl. "role"), "lato" (Engl. "side"), etc.

[12] Many algorithms of word-sense disambiguation (WSD) are dictionary and knowledge-based. These algorithms operates by means of explicit knowledge-bases since they use resources contained within machine readable dictionaries, thesauri, computational lexicons, ontologies. Algorithms of *Gloss Overlap* (Lesk,1986; Banerjee and Pedersen, 2002) belong to this approach: they base on the hypothesis that there is some kind of relation among the words that are used together within a sentence. This relation can be determined by observing for each word of the sentence all the possible definitions in a dictionary: a word is correctly disambiguated by comparing all its definitions with the definitions of the other words in the sentence and choosing the one having the higher lexical overlap.
Supervised algorithms of WSD, instead, require no access to explicit knowledge since they operate by means of statistical criteria taking into account the linguistic context of words obtained from training corpora. They base on the thesis that the local context can provide evidences for the sense disambiguation: these evidences are obtained from hand-tagged corpora, already containing information about the sense of the words and their relations. Among supervised algorithms of WSD there is the *Most Frequent Sense (MFS)* algorithm which disambiguates a word by assigning to it the most frequent sense that has been computed within the training corpus.

A lexical expression can therefore contain a certain amount of ambiguity, which enables two or more attributions to it: what, in any form, represents aspects of the language incalculability is, thus, managed with great difficulty by a machine.

Generally, problems for an automatic document processing come from the strong interaction and inter-dependence among the syntactic, semantic and pragmatic levels, which make flexible and dynamic the use of the language signs: word senses have uncertain boundaries and very often they change according to the interactions built with other elements within the contexts where they can occur and according to the extra-textual context. Therefore, to describe a document and to understand its contents it is necessary to identify not only the single signs but even the relations these signs keep up among them, firstly at a syntactic and semantic level and, secondly, at a pragmatic level, that is to say the relations the signs have with the external context and in general with the domain the document pertains to. The semantic dimension, indeed, permits to consider as acceptable only some of the possible syntactic interpretations, and the pragmatic dimension permits to solve many semantic indefiniteness.

Ambiguity can then be solved by resorting to the knowledge of both the co-text and the domain of reference where the texts is placed and used: in this sense, the domain becomes a real encyclopedia functional to the interpretation of the document sense. Not only it enables an immediate interpretation of the language signs but, considering their possible implications, it also permits further interpretations: each expression can, in fact, be subjected to a semantic interpretation and each interpretation can open to other meanings. The encyclopedic knowledge, then, provides instructions to interpret in the most complete way the document sense.

This is important above all when dealing with specialized domains which produce their own documents in their own language variety (or sublanguage): in such domains (like the bureaucratic one) the interpretation of document data is generally unique, given the technicalities introduced in the sublanguage which reduce the problems due to ambiguity and incomprehension.

4. Characteristics of specialistic languages

A specialistic language represents a sub-variety of the common and general language: it adds to the basic data of the general language specialistic data, in relation to the specificity of the concepts dealt, in order to provide the experts of the domain with a technical and rigorous terminology, so to ensure a communication without ambiguity.

Rigour and clarity represent the important characteristics of a specialistic language: the former is functional to the possibility of determining the document contents in a univocal way; the latter is functional to the possibility, for the receivers, of an easy access to these contents. Consequently, rigour and clarity depends on the terminology used in the domain: a technical word (or term) must determine its sense in the most rigorous way and convey one single meaning.

Specialized languages aim at an ideal of monosemy, that is a univocal relation between concept expressed and term designating it: each designation must exclusively represent the concept at issue.

Finally, there are unsupervised algorithms of WSD that find the correct sense of a word by computing a similarity measure among the target word and the other words within its local context. They base on the thesis that similar senses occur in similar contexts. In this case the sense of a word can be obtained from the text by *clustering* the occurrences of the word by means of these similarity measures. This process creates *lexical chains*, that are chains of words semantically linked by means of a relation of cohesion. Each occurrence of the word must belong to one and only one chain. Algorithms belonging to this approach are *Morris and Hirst's algorithm* (1991) that use a thesaurus as knowledge-base to extract the relations among terms, and *Hirst and Stonge's algorithm* (1998), that use WordNet as source for relations.

Therefore, these languages need to create their own terminology, that is to say their own set of specialistic (technical) words.

A term, or terminological unit, is the designation of a concept in a specialistic language. This designation can be:

- a word belonging to the common language that has been assigned with a new and specialistic meaning (redefinition): this is an exemplification of a specialistic re-use or even sense extension;
- a word that exclusively belongs to the specific domain (technicality): it has a univocal meaning and doesn't occur outside the domain;
- a derived form created from the basic form through the insertion of suffixes or prefixes (or both);
- a compound or a multi-word expression that has been lexicalized and forming a complete unit of sense;
- an acronym, an abbreviation, a formula;
- a loan word.

Sublanguages, then, can produce new words and expressions or assign a new and a more specialized sense to words already existing in the standard language. Operations of redefinition and technicalization, therefore, produce neologisms of sense which serve to reduce the risks coming from bad interpretations. Operations of derivation, composition and abbreviation, as well as lexicalization, can create, instead, neologisms of form that even serve to characterize the specialistic language.

A neologism can become a specialistic term of a domain only if it conveys the content of the expressed concept.

Terms, or specialistic words, are then characterized by:

- the univocality and stability of the relation established with the specific concepts they designate within the domain;
- the regular and remarkable frequency with which they are used to designate specific concepts within the documents pertaining to the domain;
- the limited combination of the structures where they occur: the most part of terms are simple and derived nouns, as well as compounds and noun phrases.

Within a specific domain, a specialistic concept can be recognized by means of:

- the set of characteristics describing it in any corpus pertaining to the domain itself;
- a definition distinguishing it from other concepts;
- a regular association with a designation.

On its side, a term is recognized by means of a regular association with a set of characteristics able to define the concept it designates. There is, therefore, a semantic stability linking the concept to the term.

As far as the structure of terms is concerned, simple terms correspond to single word (even if derived or composite), delimited from the other words by two blank spaces. But terms, as stated before, can also have a complex structure, that is to say that they can be composed of two or more words separated by blank spaces forming an expression conveying a complete and autonomous sense. Simple terms often take part into the composition of more and more complex terms designating more and more subordinate concepts. When the length makes them not much practical to use, they tend to become acronyms and to be used as simple terms, furthermore, they can be also used to compose another complex term. This is the case of a term like "RAdio Detecting And Ranging", abbreviated in RADAR and used to compose a complex expression like "meteorological radar".

These complex expressions are, then, very frequent within specialized domains, given the specificity of the matters to deal: generally these phrase structures are the output of technical uses and very often they represent specialized designations of more general concepts.

Therefore, syntagmatic relations are evidence, at a deeper level, of sense relations: words can regularly co-occurr because of their intrinsic sense which make them conceptually associated (isotopy).

It is therefore important, while analyzing a specialistic text, not to lose the overall sense of these syntagmatic sequences dispersing the single lexical items: it is necessary to process the complex term as autonomous unit of analysis. The identification of these sequences of words is then fundamental for the comprehension of the text: they obviously depend on the semantic of the text and catching them automatically is far from being simple.

Their recognition relies principally on human intervention and involves two principal steps: i) the identification of phrase structures; ii) the selection of the relevant structures designating meaningful concepts of the domain. Semi-automatic techniques in this sense are the key-word-in-context analysis, the co-occurrence analysis and the analysis of the repeated segments (Bolasco, 1999, 2004).

A central aspect for a correct document interpretation is, then, the continuous resorting to the linguistic and extra-linguistic knowledge: all texts are riddled with more or less shared knowledge, some of them are general and common, others depends on our encyclopedia, which works as a hypothesis regulating the interpretation according to the domain of use.

Thus, it is possible to state that the comprehension of specialistic documents causes: i) less problems than the comprehension of more general texts since, being more rigorous, they reduce semantic ambiguity; ii) more problems of comprehension for people who are not expert of the domain.

Conclusions

The new era of Computer Technology has brought to a dematerialization which has turned documents into automatically processable objects. Documents have acquired a new corporeity made of *bit* which can be processed with great effectiveness, modified, recomposed, decoded, transformed into sequences of words, images or sounds. A Document coincides now with a virtual and multi-sequential entity, which is open and modifiable: innovative characteristics of documents are, in fact, computability, ductility, multi-modality and content stratification.

Furthermore, the daily need to access to structured or semi-structured semantic contents within huge document collections in natural language has given an impulse to the development of ICT solutions at a methodological, technological and architectural level to serve as support for the automatic document management, and principally aimed at a semantic retrieval of the information in a logic of intelligent knowledge management.

We have proposed a general model of document and we seen how activities of text analysis are required in order to extract relevant information from them, this last one made possible by means of linguistic and statistic techniques, despite the several problems hampering the interpretation of natural language.

References

Banerjee S., Pedersen T. (2002), *An adapted Lesk algorithm for word sense disambiguation using WordNet*. In: Proceeding of the Third International Conference on Intelligent Text Processing and Computational Linguistics, Città del Messico.

Bolasco S. (1999), *Analisi multidimensionale dei dati. Metodi, strategie e criteri di interpretazione,* Roma, Carocci

Bolasco S. (2004), *L'analisi statistica dei dati testuali: intrecci problematici e prospettive.* In: Bolasco S., Cutillo E. A. Applicazioni di analisi statistica dei dati testuali, Roma, Casa Editrice Università La Sapienza

Brill E. (1995), *Unsupervised Learning of Disambiguation Rules for Part Of Speech Tagging*. In: Proceedings of the Third Workshop on Very Large Corpora.

Brill E., Wu J. (1998), *Classifier Combination for Improved Lexical Disambiguation*. In Proceedings of the Joint 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL). Montreal, Canada.

Carlberger J., Kann V. (1999), *Implementing an Efficient Part-of-Speech Tagger*. Software - Practice and Experience, 29(9): 815-832.

Carnelutti F. (1975), *Documento (teoria moderna*). In Novissimo Digesto Italiano.

Charniak E., Hendrickson C., Jacobson N., Perkowitz M. (1993), *Equations for Part of Speech Tagging*. In Proceeding of the Eleventh National Conference on Artificial Intelligence, pp. 784-789.

Chiari I. (2007), *Introduzione alla linguistica computazionale,* Roma-Bari, Editori Laterza

Daelemans W., Zavrel J. (1996), *MBT: A Memory-Based Part Of Speech Tagger Generator*. In: Proceedings of the Fourth Workshop on Very Large Corpora, pp.14-27, ACL SIGDAT.

De Mauro T. (1961), *Statistica Linguistica* In: Enciclopedia Italiana, Appendice III, Vol. II, Ist. Enciclopedia Italiana

De Mauro T. (1994), *Capire le parole,* Bari, Laterza

De Mauro T. (1997), *Guida all'uso delle parole,* Roma, Editori Riuniti

De Mauro T., Mancini F., Vedovelli M., Voghera M. (1993), *Lessico di frequenza dell'italiano parlato*, Etaslibri, Roma.

Delmonte R. (1997), *Rappresentazioni lessicali e linguistica computazionale*. In: Atti SLI, Lessico e Grammatica – Teorie Linguistiche e applicazioni lessicografiche, Roma, Bulzoni, pp. 431-462.

Dermatas E., Kokkinakis G. (1995), *Automatic stochastic tagging of natural language texts*. In: Computational Linguistics, 21:137–163.

Derose S.J. (1988), *Grammatical Category Disambiguation by Statistical Optimization*. In: Computational Linguistics, 14(1): 31-39.

Eco Umberto (1975), *Trattato di semiotica generale*. Bompiani, Milano.

Eco Umberto (1979), *Lector in fabula. La cooperazione interpretativa nei testi narrativi.* Bompiani, Milano.

Eco Umberto (1990), *I limiti dell'interpretazione.* Bompiani, Milano.

Gallino L. (1991), *Informatica e scienze umane: lo stato dell'arte,* Milano, Franco Angeli

Gigliozzi G. (1997), *il testo e il computer: manuale di informatica per gli studi letterali,* Milano, Bruno Mondadori

Hirst G., StOnge D. (1998), *Lexical chains as representations of context for the detection and correction of malatropisms WordNet: An electronical lexical database.* C.Fellbaum (editor), Cambrige, MA: The MIT Press.

Kupiec J. (1992), *Robust part-of-speech tagging using a Hidden Markov Model.* In: Compututational Speech Language, 6:225–242.

La Torre M. (2005), *Le parole che contano. Proposte di analisi testuale automatizzata,* Milano, Franco Angeli

Lenci A., Montemagni S., Pirrelli V. (2005), *Testo e computer. Introduzione alla linguistica computazionale,* Roma, Carocci Editore

Lesk M. (1986), *Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone*. In Proceedings of the 1986 SIGDOC Conference, Toronto, Canada, 24-26.

Marquez L., Rodriguez H.(1997), *Automatically Acquiring a Language Model for POS Tagging using Decision Tree*. In Proceedings of the Second Conference on Recent Advances in Natural Language Processing, RANLP '97.

Morris J., Hirst G. (1991), *Lexical cohesion computed by thesaural relations as anindicator of the structure of text*. Computational Linguistics, 18:21-45.

Orlandi T. (1990), *Informatica Umanistica,* Roma, NIS

Rizzi A. (1992), *Orientamenti attuali della statistica linguistica.*, In: *Statistica,* n. 4

Schmid H. (1994), *Probabilistic Part-of-Speech Tagging Using Decision Trees*. In Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK.

Spina S. (2001), *Fare i conti con le parole. Introduzione alla linguistica dei corpora,* Perugia, Guerra.

Tamburini F. (2000), "Annotazione grammaticale e lemmatizzazione di corpora in italiano". In Rossini Favretti (a cura di), *Linguistica e informatica*. Roma, Bulzoni, pp. 57-73.

Voutilainen A. (1995), *A syntax-based part-of-speech analyzer*. 7th European Conference Chapter Assoc. Comp. Linguistics, pages 157–164. ACL.