Infopragmatics: an efficient method for information retrieval

Ibarra Rafael and Ballesteros Silvia
National Autonomous University of Mexico.

Abstract: Based on a linguistic algorithm, supported by an uncertainty theorem, the *infopragmatics* is a new method that offers an efficient solution, but not limited, to those Spanish speaking users who try to get the most useful information from academic databases which contents is in English. Presents a search analysis, an application of the language of levels understanding table, brief considerations on the ambiguity of the term "relevance" and statistical reasons to put *infopragmatics* into action at our National University Library System.

……
Keywords: infopragmatics – information retrieval (IR) – algorithm – uncertainty theorem – pragmatics – linguistics – Spanish speakers.
……

1. Introduction

Information Retrieval is an unfinished challenge faced by today's librarians in order to give satisfaction to their users with pertinent information. There are several reasons that make this problem unsolvable, among others: information systems chaotic interfaces; users multiple categories and the unavoidable mismatch between controlled and natural vocabulary. These three critical aspects and their underlying grounds lead users in general to experience an uncertainty phase that overwhelms and inhibits them to get the information they need from information systems (ISs) also known as data bases (DB).

Furthermore, though there have been several alternatives given by ISs providers and uncountable researchers' theoretical contributions, users' demands for a suitable method to their information satisfaction have not been decreased sufficiently: they are still looking for an appropriate method. *Infopragmatics* is offered to relieve those crucial aspects by means of linguistic tools: a *linguistic algorithm*, an *uncertainty theorem* and a *linguistic storm* (*lingstorm*) – that result in an effective approach to get an efficient non-robotic alternative to satisfy users' information demands, specially for, but not limited to, those Non Native English Speakers (NNES).

Before going into deep, it should be mentioned that though there is a variety of DB that provide data in several formats: text, image, sound, video, and their possible combinations, the present paper will only consider commercial ISs companies with academic information, such as the ones offered by ProQuest, EBSCO and Elsevier because they are the ones used by the involved community. Google, Yahoo and the like are not considered in this study due to their nature as search engines, commercial sponsored interests and aims, though they are information alternatives many user take.

Information systems

The commercial ISs above mentioned present chaotic interfaces search pages

and a quick access cumbersome help key (F1) that clearly show how unsuitable they might be for the numerous users' kinds, but especially for a novice user whose language is not English. For example, DB interfaces present more than 25 elements that are non essential for a novice user. Some of such elements are: *Pay Per View, Alerts, Favourite Journals, Quick Links, External Links, My settings, Number of pages, Cover story, SmartText searching, Also search within the full text of the articles, New features,* and several other elements which are definitively out of the novice user's interest. The rest of elements – ten - are more appropriate for mid or expert users.

According to the last reported inform, the top three consulted databases at the UNAM in 2006 were *Academic Search Premier* (25,723 consultations), *Elsevier Science* (7753 consultations), *PsychINFO* (5472 consultations). After reviewing the interfaces of each of these providers, it was found that they fit adequately for a certified librarian or user by the companies, as well as for a robot to read and select the check lists presented in an instant, but they are definitively not easy to deal with for a standard librarian or user whose language is Spanish or some other language different than English.

The lack of uniformity and troublesome interfaces of the ISs prevent several users to use them and prefer the search engines as Google Scholar that offers a suitable interface of no more than nine elements in English and five in Spanish language, which is a key component that is dealt in the following section.

2. Spanish Speakers Experience with ISs in English language

The National Autonomous University of Mexico Library System (UNAM LS) offers access to more than 198 databases to its users concerning their fields of study and 91% of them are in English language. Its services are given in more than 140 libraries, but the analysis presented and referred in this work took place at the Central Library (CB) where the interaction between a librarian and a user is both: spoken and in Spanish. Nevertheless, the interaction between humans and the ISs must be written and in English, which already sets a series of probable problems: bad spelling, wrong affixation, use of natural language in opposition to the controlled vocabulary and a big and chaotic number of the DB tools that come within as those alluded previously.

Every year more than 9,000 graduate students and more than 37,000 undergraduate enrol as students[1]. Despite the fact that many librarians hold their posts during several years and offer annual courses related to databases use, thousands of users are mostly new to ISs.

2.1 Kinds of users and their language

Users are unique in their ideas, thoughts, and needs, so the way in which they express their information queries is as different as can be. On the other side, Spanish speaker users, either students or librarians, are not one of a kind and internationally share characteristics that can be summed up in the following, but partial categories referred by Stubinz and Whighly (2003): Pip, the impatient; Odysseus, the dogged; Ishmael, the exploratory; Hamlet, the confused; Ophelia, the deranged and Don Quixote, the idealist. Though this is a limited variety of users, it can be appreciated that, besides the human features, users speak

[1] http://www.estadistica.unam.mx/agenda/agendas/2008/disco/xls/124.xls

different languages and adapted to modern times, their languages would be: English, Greek, Hebrew, Danish and Spanish.

With this scope, it is not difficult to imagine the effects that the merge between these kinds of users, different languages and commercial ISs' chaotic interfaces may cause. In current times it is evident that researchers, teachers at all levels have at hand a great number of strategies with numerous modifications and concerns to know, manage and improve the manners and channels of user services; Katsirikou and Skiadas, (2001) list 23 processing actions that comprise the opening and the closing dialog in an information request and that go from finding the appropriate electronic resource to indirectly and unwittingly provide personal information on one's activities, no matter the language.

2.2 Controlled vocabulary and Natural language

One of the oldest antecedents of controlled vocabulary and natural language is the dispute between two groups: the *anomalists* and the *analogists*. The first group stated that human language should be spoken and lived in the way in which it is articulated; and the second one held that humans should express their ideas in an appropriate manner. As the years passed, the groups turned to be prescriptive and descriptive.

This dispute has not finished yet, as it can clearly be appreciated in modern classrooms when users are using their natural language in order to find controlled vocabulary information for their formal research. They are not aware that the semantic cracks that exist between ISs and them (Ibarra, 2007) is, in most cases, the reason why after a fruitless information search they end feeling uncertain and frustrated. They would certainly prefer feeling "found" than feeling "lost" if they knew how. An option would be to handle the theorem of uncertainty presented in here.

3. Infopragmatics

Infopragmatics is an efficient method to get pertinent information form ISs. It is based on a linguistic perspective aimed to those users, whose language is not English, paying particular focus but not limited to – Spanish speakers. Infopragmatics may be of highly considerable help for those academicians who can not get pertinent information from commercial ISs such as the ones provided by Elsevier, Ebsco, ProQuest and the like and it consists of an uncertainty theorem, a linguistic algorithm and a linguistic storm (lingstorm).

3.1 The uncertainty theorem

When users can not get the information they need from a DB, either guided by librarians or by using the quick guides given by ISs providers, they wonder why, and there is not a chance to give them a clue to reveal what the problems were, since they had the chance to read *relevant* displayed records from a scientific ISs. For example, Jaffe (1988: 759) points out that after a test applied to ISs users without more help than the one within DB, students were thwarted by Boolean, however. They sought the quick "hits" possible in an index-controlled database, and they were frequently confused when non-relevant citations appeared in a Boolean search, even though all the search terms were represented in the citation. After long reading sessions users got tired and fall into a state of uncertainty. Ibarra (1999) presented an approach considering diverse linguistic theories from an ethnolinguistic perspective and several analyses based on 120

real information searches with similar results.

So far, ISs do not offer any aid to keep the users away from the resulted perplexity; a gap which is expected to be covered by the following uncertainty theorem as an efficient contrivance to steer users to step more confidently in their information goals:

I.- If there is no information is due to...

User lacks lexical abilities
Syntax is wrong
Used terms are not equivalent in meaning
Context is no the appropriate
There is nothing in the used data base

II.- If the information doles not fit is due to. . .

Inappropriate technical language
Incorrect spelling
Key words or descriptors are not such

To reinforce the previous considerations, it is convenient to illustrate the diverse and synthesised linguistic possibilities that make Spanish native speakers fall apart from their information goals at the following understanding levels Liddy (2007): *Morphological*, bad spelling; *Lexical*, stating their needs; *Syntactic*, grammar rules; *Semantic*, opposition between the natural language and controlled vocabulary; *Discourse*, the way in which librarian and user interact to offer and receive ISs information respectively and, *Pragmatic*, the set of strategies that allow the users/librarians contrive their information needs. The next figure represents an information request on *Bolsa Mexicana de Valores*.

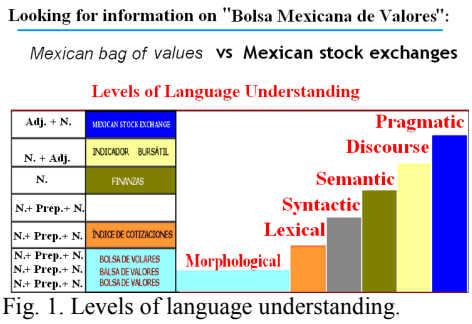

Fig. 1. Levels of language understanding.

It is evident the bad use of English language, but it is not possible to deny that "bag" means "bolsa", "Mexican" mexicana and "values" "valores"; or *sac de valeurs* vs. *Bourse* . Though this information query was taken several years ago, it was made the same trial on April 19, 2009, it was found that Academic Search Premier (Ebsco): and Science Direct (Elsevier), presented *relevant* results: 365540 and 1475 records respectively.

Users may not know much about levels of language and it is common that they follow a critical route: The first step is the lexical confusion, lack of technical dictionaries, indices, subject headings or thesaurus and misspelling. The second step is facing the ISs chaotic interface and dim help. The third step is reading the displayed results and after, selecting and printing or e-mailing undesired records, to justify, in most cases, the usage of ISs. Each of the previously identified errors may cause a variety of problems not just for Spanish Native speakers, but for all those native speakers whose language is not English.

Opportunely, there is another tool that identifies the critical route and suggests and ideal one, besides indicating the steps to be taken to warrant a practical solution to that users' urgent demand to get pertinent records from ISs.

3.2 Relevance

One striking and screened fact that in some way contributes to users' uncertainty is the term *relevance*. On the one hand, the results presented by the DB are offered based on the *relevance* they represent from the ISs, which means, from the engineering point of view, in several cases, the number of times that a term appears on a document.

On the other hand, entering to the definitions of a sense brings within a series of tasks that has come full circle, returning most recently to empirical methods and corpus-based analyses that characterize some of the earliest attempts to solve the problem. Among the numerous efforts that have been done Ide and Véronis (1998) remark AI-base methods, Symbolic methods, connectionist methods, knowledge based-methods, machine readable dictionaries, computational lexicons, corpus based methods (automatic sense-tagging and overcoming data sparseness.. However, the term *relevance* used by the ISs providers to present the results of their Information Retrieval Systems unwillingly masks the meaning *repetition* and *anaphora,* so users tend to follow a critical route shown in the next section.

3.3 Linguistic Algorithm

The linguistic algorithm is the illustration of the series of linguistic operations to resolve the problems intrinsic to IR. It can be used as a guiding map to make users aware of the pair of routes within and, alternatively, considering the use of quick actions to satisfy their information needs in opposition to the critical route which issues were discussed previously.

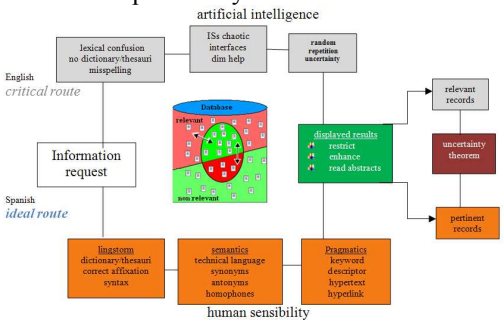

Fig. 2. Linguistic algorithm.

The first element of the ideal route comprises a quick action that requires using linguistic sources, either in print or electronic versions, to ensure the following steps. Specialised dictionary and thesaurus will allow the user to approach the *lingstorm*, "linguistic storm" (Fig. 3), to be used in a similar way as a brainstorm works in a group or individual creativity exercise, a constant along the IR, needed to leap from natural language to controlled vocabulary; the correct affixation deciphers the differences among verbs, nouns and adverbs that commonly cause semantic breaks when users shift from Spanish to English; Syntax determines the results in noun phrases; technical language clinches users´ familiarity with controlled vocabulary; Synonyms enhance semantic horizons and include spelling varieties when dealing with non Latin characters;

though antonyms and homophones take place during verbal interaction, they fathom a temporary aphasia; key words and descriptors are easily found along the records´ reading and can be retrieved to set alternative search strategy; hypertext and hyperlinks are options that are attractive to follow as soon as they appear on the information records in order to avoid semantic unsettlement.

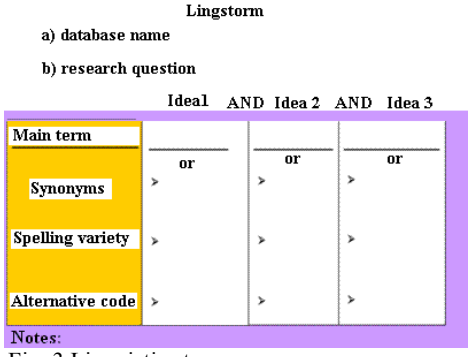

Fig. 3 Linguistic storm.

3.4 Statistical reasons and technical implementation of the *infopragmatics* in the National University Library System

The technical implementation will demand observing the number of users, the collections and two users' scenarios: distance and On-site. As it was mentioned before, the UNAM received 46,000 - undergraduate and graduate – potential ISs users; according to the last statistical data, the UNAM's Library System, up to April, 2009, there is more than 2 million of documents, comprising books, thesis, ISs, serials, printed and electronic.

More specifically, the available 150 ISs – in English - comprise access to more than 400,000,000 of records, besides access to more than 1,000,000 of journals' issues. On the other hand, the information requests in ISs daily average is 20,000.

Though it is not known the exact number of those users who do not receive librarian personal assistance, and the regular schedules they utilise the ISs, it is known that not all users have time or willing to receive instructions on the infopragmatics method.

Considering the previous points, besides the level and kind of users, it is being designed a pair of brief workshops: On-site and distance. The first one consists of three parts: 15 minutes to explain the infopragmatics; 15 minutes of practice and; 15 for evaluation and feedback. The second workshop – distance- would be available on line and it is the users themselves who must set their times to profit from it, though there will be suggested steps and a feedback section via e-mail. Currently, there is a quick guide to "Retrieve pertinent information in English" in the Digital Library main web page <bidi.unam.mx>.
From a qualitative point of view, infopragmatics workshops would be more effective in the On-site mode, nevertheless, due to the large quantity of users, and considering the three single steps of the algorithm, a quick guide was designed and entitled "Recupera información pertinente en inglés" (retrieve pertinent information in English) which is available at http://132.248.9.9:8084/infopragmatica/.

It is necessary to mention that this quick guide includes hyperlinks to electronic translators, specialised dictionaries and thesauri, which quality was evaluated and later, catalogued in the UNAM´s Digital Library.

There is a hyperlink that allows users to express their opinions on the utilisation of the quick guide at the bottom of the page. With the obtained results, that are stored in a database, the design would be improved.

……

4. Conclusions

Information retrieval has a wide variety of problems that can only be solved by appropriate solutions, that is, if there are language disagreements, the clarifications must be given in language agreements. In this paper the common linguistic difficulties that affect Spanish speakers were identified as well as their corresponding explanations. Infopragmatics is an agile method that gives acute dynamism that allows ISs users to get pertinent information in opposition to the *relevant* ones that result from a critical route. The linguistic algorithm presented in here represents an accomplished tool that can be easily adapted to any language, included English, to serve as an information productive device.

……

References

Ibarra, R., (1999). *Aprovechamiento y optimización de los recursos tecnológicos en la búsqueda y recuperación de información en CD-ROM basados en estrategias lingüísticas.* **Masters' dissertation on Applied Linguistics. UNAM.**

Ibarra, R., (2009): Algunas grietas semánticas en la recuperación de información: una perspectiva deconstructiva para una solución pragmática. Proceedings of the *I Simposio Internacional sobre Organización del Conocimiento: Bibliotecología y Terminología*, **305-320.**

Ide, N. and Véronis, J., (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24, 1, 1-40.

Jaffe, J. G., (1988), **For Undergraduates: InfoTrac MAGAZINE INDEX Plus or WILSONDISC with Reader's Guide and Humanities Index?** *American Libraries*, Vol. 19, No. 9, 759-61.

Katsirikou, A. and Skiadas C. H., (2001). Chaos in the library environment, *Library Management*, Vol.22, No. 6/7, 278 – 287.

Liddy, Elizabeth D., (2007). Whither Come the Words? Presented at the *CENDI Subject Analysis and Retrieval Working Group Conference:* **Controlled Vocabulary and the Internet.** [On line].

Stubinz. J. and Whighly, S., (c. 1994) Information Retrieval System Design for Very High Effectiveness. [On line].