Going beyond simple keyword search in the next generation of Information Search Tools

Anastasio Molano
Denodo Technologies Inc.
Almirante Francisco Moreno, 5
28040 Madrid - Spain
amolano@denodo.com

Index

Introduction

Wouldn't be nice if you could receive an exact answer when you query a search engine, instead of a list of URL's? Questions such as "What is an iceberg" or "What is the distance between Rome and Paris?" would receive a precise answer, rather than a list of related documents. This will be possible in the short future, thanks to the evolution of Natural Language Processing techniques (NLP in short).

Nowadays the volume of information in digital format on the Internet and corporate Intranets has increased to such an extent, that there is a growing need for tools that help people to locate, filter and manage these resources in an efficient and optimal way.

Recent advances in Human Language Technologies has fostered the outcome of a new generation of search tools which make use of Natural Language Processing techniques and resources to improve its search capabilities.

Search tools have been traditionally based on classical Information Retrieval techniques, for example some kind of Boolean search or probabilistic retrieval method. In these systems search does not take into account the underlying linguistic properties of text.

NLP technologies go beyond traditional Information Retrieval techniques enabling a system to accomplish a human-like understanding of text, and thus, permitting to extract useful meaning from unstructured text.

Search companies such as Ask Jeeves, Convera, Northern Light, Verity, SmartLogik, Q-Go, and Cognit among others, have incorporated NLP techniques in their search solutions.

Expectations are high, as these tools are having a great impact on the industry, especially on large companies corporate Intranet searchers, and generally, in those applications in which searching efficiently over large document repositories is crucial (e.g. Digital Libraries, Medicine databases, Legal databases, Competitive Intelligence tools, etc.). The current relevance of multilingual, cross language and interactive retrieval will further increase demand on this kind of technologies.

Given the size of digital information universally available today, along searching itself, other complementary information processing types are required, where these techniques are finding its niche, such as automatic text categorization, filtering and summarization.

We can identify the following key search applications that exploit NLP techniques:

- Information retrieval.
- Multilingual, cross-language retrieval.
- Question answering.
- Document categorization.
- Document summarization.
- Text Mining.
- Information Extraction (including handling of XML documents).

Let's review these techniques at the first place.

Language Technologies

NLP comprises those theories and technologies, which enable a system to exploit linguistic properties of text in order to extract meaning from it. Understanding word meanings and their association with other words within a sentence structure is key to understand the true meaning of text.

Linguistic knowledge includes morphological, syntactic and semantic information that can be applied within the information retrieval process to, for example, expand queries with related terms (e.g. synonyms) and thus retrieving a larger amount of relevant documents.

Lexical Resources

NLP techniques for information retrieval relies on the use of lexical resources, being the most common ones *Machine-Readable Dictionaries - MRD* (e.g. Longman's Dictionary of Contemporary English, LDOCE), inventories of words with concise description of meanings and some morphological and syntactic information, and *Thesauri* (e.g. Roget's), which organize words on the basis of their meanings (rather than alphabetically).

An even richer resource than a MRD or a Thesaurus is a *Lexical Knowledge Base*, a fully structured computational lexicon where word forms are associated according to morphosyntactic, semantic and other kinds of information.

There has been a big effort to build comprehensive lexical knowledge data bases during the last decade, both in the USA and Europe.

The Cognitive Science Laboratory at Princeton University developed WordNet at 1995 (currently WordNet 1.7.1), a large-scale, domain independent, freely available lexical knowledge base for the English language, where information is organized around logical grouping of related terms called *synsets* (or synonym sets), each of which consists of a list of synonymous word forms and semantic pointers that describe relationships between the current synset and other synsets. The semantic content and the large coverage of WordNet make it a powerful tool to perform conceptual text retrieval.

The European counterpart is EuroWordNet, developed under an EC funded project within the Telematics Applications Programme, which finished at 1999. EuroWordNet is a multilingual database, which includes semantic relations between words for Dutch, Italian, Spanish, German, French, Czech and Estonian. Within EuroWordNet an Inter-Lingual-Index was created to interconnect the languages in such a way that it is possible to go from the words in one language to similar words in any other language, an interesting feature that permits conceptual and cross-language information retrieval as we will see below.

Other similar lexical knowledge bases are currently being developed for Swedish, Norway, Danish, Greek, Portuguese, Basque, Catalan, Romanian, Lithuan, Russian, Bulgarian and Slovenic.

These multilingual databases constitute a highly valuable resource to be exploited by searchers to perform multilingual cross-language text retrieval.

NLP techniques

Mono-lingual text retrieval

NLP techniques can be used at all the stages of the information retrieval process:

- At indexing time, we can make use of morphological analysis, Part-Of-Speech tagging, syntactic analysis and finally semantic analysis.

- At querying time, queries are indexed following the same techniques, additionally lexical resources can be used to expand the query with related terms.

Let's explain these issues in more detail.

The indexing process starts with removal of stop words from original text, followed by *stemming*, reduction of words to some base form.

NLP based stemming, also known as *lemmatization*, applies morphological analysis to extract the base form of a word, and checks base forms against a Machine Readable Dictionary (e.g. LDOCE for the English language), assuring real word stems (e.g. a search for "go" can be extended to include a search for "went", a classical stemmer would only identify "go" and "going", "gone", "goes", etc.).

Part-Of-Speech taggers assigns part of speech tags to words reflecting their syntactic category (e.g. noun, adjective, verb, adverb, etc.). More advanced taggers attempt to recognize proper names, acronyms, phrasal constructions, etc., as single tokens, for example, "New York" would be viewed as a single unit rather than just as a sequence of two words in the text.

Syntactic analysis processes each sentence to build a tree structure of phrases comprising nouns, verbs, prepositions and conjunctions. Once this portion of sentence analysis is completed, the semantic analysis can proceed to synthesize

these multi-word structures into meaningful concept relationships.

At querying time, NLP techniques can be applied to either expand the query with semantically related terms (e.g. with synonyms of the relevant words in the query, taken from a Lexical Knowledge Base such as WordNet), or to help comparing queries against documents to improve search precision.

Multilingual cross-language retrieval

Multilingual cross-language retrieval refers to retrieval of documents in different languages regardless of the language in which the query is performed (e.g. we could perform a query in English and receive a relevant document written in Spanish). EuroWordNet has been successfully used to implement this kind of multilingual search for several European languages.

Question Answering

Question Answering is a new breed of search applications in which the user just prompts a query expecting to receive an appropriate and exact answer to it, instead of a list of documents that may contain the answer. "When did Hawaii become a state?" or "How far is it from Denver to Aspen?", "What is an iceberg?", are questions that could be answered with such systems.

This is a very interesting application field of NLP techniques that has received much attention in the Text Retrieval Conference (TREC) with a specific track for it.

Market situation and prospects

Internet Searchers

Traditionally, companies that operate Internet crawlers have been reluctant to apply NLP techniques in their search solutions, mainly due to scalability problems derived of its high computing demands. However, computing power has caught up to the level of requirements of complex NLP systems, so it is expected that assuming a typical crawler infrastructure based on a networked set of powerful inexpensive computers, this will no be a problem in the future, so we will see crawlers applying advanced NLP techniques as a common feature very soon.

Notwithstanding, most of existing search engines make use of at least simple NLP techniques, in order to interpret queries written in question form, the so called *natural language queries*, and to regularize singular an plural form of words in the index.

Some Internet searchers already feature advanced NLP techniques. For example, Northern Light employs NLP to categorize documents providing specialized information folders, which are used to improve relevance ranking, and to filter and organize the results. NLP is at the core of Ask Jeeves searcher, which applies grammar processing, tokenization, stemming, stop-wording, parsing and semantic analysis in its search process.

The Norwegian company FAST Search & Transfer, provider of search infrastructure for Alltheweb and Lycos, has featured a package for *advanced linguistic*, which includes lemmatization, approximate match, and phrasing, the ability to identify phrases in sentences. This package is available for English and Spanish.

Still there is a long way to go, as very few searchers make use of already available Lexical Knowledge Bases such as WordNet, or more importantly, multilingual counterparts such as EuroWordNet.
GRAF. 1(see the document NLP_Examples)
Additionally, *Question Answering* seems to be a very promising approach for a new age of more accurate and efficient search engines. At the present time, no commercial Internet searcher has included this feature in its core technology, however demand is there, as recently launched Google's Answers Service demonstrates.

Corporate Intranet Searchers

The bulk of information managed in large corporations is nowadays in electronic format. Within a large corporation there is a huge amount of documents in dispersed repositories, Intranet Web pages, e-mails, etc., which are critical for most of the business operations of the company (e.g. product information, client historical, financial information, etc.).

Most of large corporations, such as banks, Telecoms, utilities, have already incorporated, or are in the process of incorporating a *Intranet corporate searcher*. NLP-empowered searchers are a competitive alternative to other technologies for this purpose, as they can successfully satisfy the following usual requirements:

- Precise document retrieval.
- Automatic document categorization and summarization.
- Multilingual retrieval (usual need within international corporations).

Among search companies that focus on the corporate domain and make use of NLP techniques, the most representative ones are Convera and SmartLogik.

Convera's RetrievalWare performs natural language processing (morphology analysis and linguistic pattern matching), and search term expansion of queries with a proprietary lexicon, the RetrievalWare Semantic Network, a collection of approximately 500,000 English words including semantic relations among them. Also it includes a powerful tool for categorization.

SmartLogik includes stemmers for 10 languages, and it uses a thesaurus for query term expansion. It features both a searcher and a categorizer.

Verity, the main player in the market of Intranet search, has extended its search technology to support NLP. It makes use of stemming and synonym expansion (a thesauri is included for each of the supported languages). The system administrator can select lexical and grammatical analysis among the available search options.

Inktomi offers natural language queries across XML data repositories, avoiding the use of more powerful but complex SQL queries, as information mediators like Denodo support.

IDC predicts that in the long term search will be embedded in most enterprise applications, as a function of the portal, infrastructure or gateway, even in small and medium size enterprises (SME's), so demand for these technologies in the corporate domain can be very high. That´s the reason why traditional Internet searchers such as Google, Northern Light , FAST or Inktomi are now moving their business strategy towards this very appealing market.

Other emerging application fields

Medical institutions require powerful searching tools to explore large repositories of information about diseases (scientific articles, studies, etc.), pharmaceutical products, etc. These search applications require a very high precision and consequently, simple keyword based search tools are not suitable. Also, summarization is a usual requirement in this domain to quickly visualize document contents. NLP tools provide all these needs. Worth to mention is the IST LIQUID project, which aims at developing a cross-lingual information retrieval system in the field of gastroenterology.

We can find similar information retrieval requirements on legal databases and public digital libraries. The HERMES project, funded by the Spanish Ministry of Science and Technology, focuses on applications to facilitate the retrieval of multilingual textual information in the field of Digital Libraries. Automatic categorization, summarization and concept extraction based on linguistic engineering will be implemented in the project.

An interesting application field of NLP tools is *Competitive Intelligence*, a discipline that is getting more and more importance in the business world, particularly for technology companies.

Competitive Intelligence tools perform continuous monitoring of a large amount of information sources, searching for new advances, patents, products, relevant legal and administrative information, etc., which might be crucial for companies, specially for SME's, to make successful business decisions. NLP tools such as KnowledgeGist from Invention Machine are having a great impact in this field.

European initiatives and market prospects

European NLP industry is still very young, although there are some important players such as Albert, and the Xerox Research Center Europe. Other start-up companies are appearing in the market, such as the Norwegian Cognit and the Dutch company Irion.

Among EC initiatives to fuel the European NLP industry, worth to mention is the Cross-Language Evaluation Forum (CLEF), funded under the IST Programme, which has been recently created to promote monolingual and cross-language retrieval in European languages, as the European initiative parallel to TREC.

The goal of this Forum is to promote the development of European cross-language retrieval systems, to guarantee European competitiveness in the global marketplace.

Markets studies for Content Management (CM) and Knowledge Management (KM), areas where information retrieval tools can be located, reveals a positive trend in the spend on these kind of technologies.

As the Strategy Partners "Content Management Europe 2001-2003" European market report states, content management is the fastest growing IT sector, while much of the IT industry is in recession. *Information Retrieval* tools, and consequently those based on NLP-techniques are included within this umbrella, along with electronic document management (EDM), enterprise information portals (EIPs), electronic publishing and collaborative filtering.

The worldwide spend on CM was $4,766 millions in 2001, and it is forecast to grow at a rate of 29.5% in the period 2001-2003, to reach a total market of $10,445 millions in 2003.

The spend in Europe was of $1,315 millions in 2001, with a expected growth rate of 34.5% in the period 2001-2003

(higher than worldwide), to arrive at a total market of $3,325 millions in 2003.

As we can see from Strategy Partners forecast, CM market will grow at a high rate during the 2001-2003 period, specially in Europe, and NLP tools can play an important role on this.

Market studies for Knowledge Management exhibit a similar behavior, with an estimated worldwide spend in KM software and services of $6,000 millions in 2002, and a forecast of $14,000 millions in 2006 (source: IDC). Search tools including those based on NLP are at the heart of this movement, so perspectives for the future are very good.

Research in Spain and market prospects

In Spain, a multicultural country with four official languages (Spanish, Basque, Galician and Catalan), there are a healthy number of important NLP research groups in the field of Information Retrieval at the University of Alicante (with important presence in TREC with Question-Answering applications), UNED, Technical University of Catalonia (UPC - TALP research Centre) and the University of Barcelona all of them involved in EuroWordNet project, University of Basque Country (Ixa research group involved in extending EuroWordNet with Basque language), and the Technical University of Madrid (UPM).
GRAF.2 (See attached file NLP_Examples)
Additionally there are some emerging start-ups companies such as Daedalus, launched by researchers from the Technical University of Madrid, and CLiC, start-up from the University of Barcelona.

Spain has a strong position in the Content Management market (source: Strategy Partner), as Telecom and media companies, early adopters of this kind of solutions, have a strong presence in the country, so we can anticipate good prospects for the expansion of NLP tools as well.

Iberoamerican initiatives

Spanish and Portuguese languages presence on Internet, largely dominated by English language today, depends on the development of NLP techniques and resources for both languages. They are an important concern that Iberoamerican CYTED Programme has taken into account through RITOS2 Research Network, which includes NLP research groups from 14 Iberoamerican countries.

Conclusions

In conclusion, we can advance that Natural Language Processing techniques will play an important role in the Knowledge Economy in the years to come, specially in Europe, with very important applications fields such as the ones described in this article.

LINKS OF INTEREST

Information Retrieval Evaluation Forums:

TREC: http://trec.nist.gov

Cross-Language Evaluation Forum: http://www.clef - campaign.com

Lexical Resources:

Wordnet: http://www.cogsci.princeton.edu/~wn/

EuroWordNet: www.illc.uva.nl/EuroWordNet/

Tutorials:

Language Engineering Open Distance Course: http://rayuela.ieec.uned.es/~ircourse/

Search companies:

http://www.northernlight.com

http://www.ask.com

http://www.google.com

http://www.convera.com

http://www.smartlogik.com

http://www.q-go.com

http://www.verity.com

http://www.albert.com

http://www.cognit.com

http://www.xerox.com

http://www.irion.nl

http://www.invention-machine.com

http://www.denodo.com

Market reports and prospects:

http://www.strategy-partners.com

http://www.infonortics.com/searchengines/index.html

The Future of Search. Clare Hart. Search Engine Meeting April 2002. (http://infonortics.com/searchengines/sh02/02slides/hart.pdf)

http://www.gartner.com

NLP Research Projects in the field of Information Retrieval

HERMES: http://terral.lsi.uned.es/hermes/index.html

LIQUID: http://liquid.sema.es

NLP Research Centers in Spain in the field of Information Retrieval:

http://www.talp.upc.es

http://sensei.lsi.uned.es/NLP/

http://dlsi.ua.es

http://ixa.si.ehu.es/Ixa

http://www.dit.upm.es

NLP start-ups in Spain:

http://clic.fil.ub.es/index.shtml

http://www.daedalus.com

Iberoamerican Research Networks:

RITOS2: http://emilia.dc.fi.udc.es/Ritos2/