Natural Language Processing in Textual Information Retrieval and Related Topics

Mari Vallez; Rafael Pedraza-Jimenez

Citación recomendada: Mari Vallez; Rafael Pedraza-Jimenez. *Natural Language Processing in Textual Information Retrieval and Related Topics* [en linea]. "Hipertext.net", num. 5, 2007. <http://www.hipertext.net>

1. Introduction

"Natural Language Processing" (NLP) as a discipline has been developing for many years. It was formed in 1960 as a sub-field of Artificial Intelligence and Linguistics, with the aim of studying problems in the automatic generation and understanding of natural language.

At first its methods were widely accepted and successful. However, when applied in controlled environments and with a generic vocabulary, many problems arose. Among those problems were polysemy and synonymy.

In recent years contributions to this field have improved substantially, allowing for the processing of huge amounts of textual information with an acceptable level of efficacy. An example of this is the application of these techniques as an essential component in web search engines, in automated translation tools or in summary generators [Baeza-Yates, 2004].

This article aims to review the main characteristics of natural language processing techniques, focusing on its application in information retrieval and related topics Specifically, in the second section we will study the different problems in automatic natural language processing; in the third section we will describe the key methodologies of NLP applied in information retrieval; and in the fourth section we will state several fields of research related to information retrieval and natural language processing; finally we present the conclusions and an annexe (Annexe 1) showing some of the particular aspects of NLP in Spanish.

2. Problems with natural language processing: linguistic variation and ambiguity

Natural language, understood as a tool that people use to express themselves, has specific properties that reduce the efficacy of textual information retrieval systems. These properties are linguistic variation and ambiguity. By linguistic variation we mean the possibility of using different words or expressions to communicate the same idea. Linguistic ambiguity is when a word or phrase allows for more than one interpretation.

Both phenomenons affect the information retrieval process, even though in different ways. Linguistic variation provokes document silence, that is, the omission of relevant documents that fulfil information needs, because the same terms were not used as those found in the document. Ambiguity, on the other hand, implies document noise, or the inclusion of non-meaningful documents, since documents were retrieved that used the same term but with a different meaning. These characteristics make automated language processing considerably difficult. The following is a set of examples that show the repercussions of these phenomena in information retrieval:

At a morphological level, the same word may play different morph-syntactic roles relative to the context in which they appear, causing ambiguity problems (example 1).

Example 1. A notebook was the present that his wife gave him when all of us were present at the party.

In this case, the word "present" acts both as an adjective and noun, and with different meanings.

At a syntactic level, focusing on the study of established relations between words to form larger linguistic units, phrases and sentences, ambiguities are produced as a consequence of the possibility of associating a sentence with more than one syntactic structure. On the other hand, this variation supposes the possibility of expressing the same idea, but changing the order of the sentence's syntactic structure. (example 2).

Example 2. He ate the chocolates on the plane.

This example could mean that "He ate the chocolates that were in the plane" or that "He ate the chocolates when he was flying in the plane."

At a semantic level we study the meaning of a word and sentence by studying the meaning of each of the words in it. Ambiguity is produced because a word can have one or various meanings, which is known as polysemy (example 3).

Example 3. Paul was reading a newspaper in the bank.

The term "bank" could refer to a financial institution or a mound.

And we must also keep in mind lexical variation which refers to the possibility of using different terms for the same meaning, that is, a synonymy (example 4):

Example 4: Car / Auto / Automobile.

At a pragmatic level, based on a language's relationship to its context, we often can not use a literal and automated interpretation of the terms used. In specific circumstances, the sense of the words in the sentence must be interpreted at a level that includes the context in which the sentence is found. (example 5).

Example 5. Give me a break.

Here we are asking for rest from work, or we could be asking the perceiver to leave us alone.

Another important topic is ambiguity provoked by an anaphora, for example, the presence of pronouns and adverbs that refer to something that was previously mentioned (example 6).

Example 6. It was terrible for him she had not to manipulate it.

Who is he? And she? What was not manipulated? It is impossible to understand this sentence out of context.

All of these examples demonstrate the complexity of language, and that any automated processing is not easy or obvious.

3. Natural Language processing in textual information retrieval

As the reader has probably already deduced, the complexity associated with natural language is especially key when retrieving textual information [Baeza-Yates, 1999] to satisfy a user's information needs. This is why in Textual Information Retrieval, NLP techniques are often used [Allan, 2000] both for facilitating descriptions of document content and for presenting the user's query, all with the aim of comparing both descriptions and presenting the user the documents that best satisfy their information needs.

In other words, a textual information retrieval system carries out the following tasks in response to a user's query (image 1):

1. Indexing the collection of documents: in this phase, NLP techniques are applied to generate an index containing document descriptions. Normally each document is described through a set of terms that, in theory, best represents its content.
2. When a user formulates a query, the system analyses it, and if necessary, transforms it with the hope of representing the user's information needs in the same way as the document content is represented.
3. The system compares the description of each document with that of the query, and presents the user with those documents whose descriptions are closest to the query description.
4. The results are usually listed in order of relevancy, that is, by the level of similarity between the document and query descriptions.

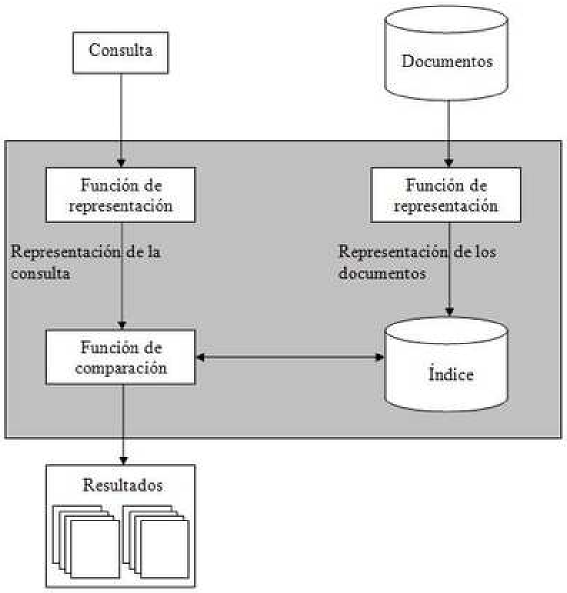

Consulta
Documentos
Función de representación
Función de representación
Representación de la consulta
Representación de los documentos
Función de comparación
Índice
Resultados

Image 1: The architecture of an information retrieval system

As of now there are no NLP techniques that allow us to extract a document's or query's meaning without any mistakes. In fact, the scientific community is divided on the procedure to follow in reaching this goal. In the following section we will explain the functions and peculiarities of the two key approaches to natural language processing: a statistical approach and a linguistic focus. Both proposals differ considerably, even though in practice natural language processing systems use a mixed approach, combining techniques from both focuses.

3.1. Statistical processing of natural language

Statistical processing of natural language [Manning, 1999] represents the classical model of information retrieval systems, and is characterised from each document's set of key words, known as the terms index.

This is a very simple focus based on the "bag of words." In this approach, all words in a document are treated as its index terms. Moreover, each term is assigned a weight in function of its importance, usually determined by its appearance frequency within the document. This way the word's order, structure, meaning, etc, are not taken into consideration.

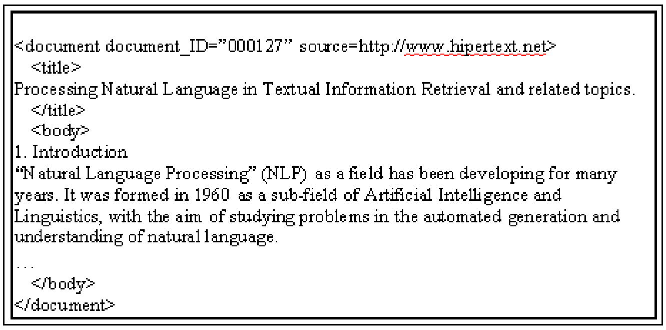
These models are then limited to pairing the documents' words with that of the query's. Its simplicity and efficacy has become the most commonly used contemporary models in textual information retrieval systems.

This document processing model involves the following stages:

a)Document pre-processing: fundamentally consisting in preparing the documents for its parameterisation, eliminating any elements considered as superfluous.

b)Parameterisation: a stage of minimal complexity once the relevant terms have been identified. This consists in quantifying the document's characteristics (that is, the terms).

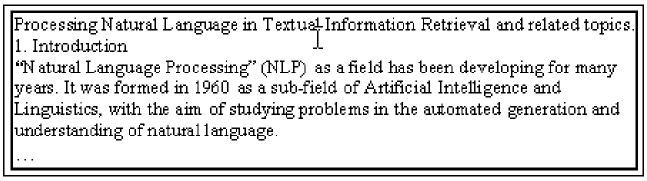
Below we will illustrate their function using this paper's first paragraph as an example, assuming that it is XML tagged. So the document on which we would apply the pre-processed and parameterisation techniques would be the following:



```
<document document_ID=”000127” source=http://www.hipertext.net>
   <title>
Processing Natural Language in Textual Information Retrieval and related topics.
   </title>
   <body>
1. Introduction
“Natural Language Processing” (NLP) as a field has been developing for many
years. It was formed in 1960 as a sub-field of Artificial Intelligence and
Linguistics, with the aim of studying problems in the automated generation and
understanding of natural language.
...
   </body>
</document>
```

Document pre-processing consists of three basic phases:

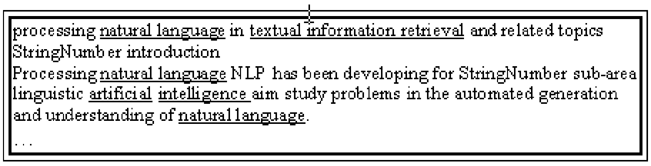
1. Elimination of the elements in the document that are not for indexing (stripping), such as some document tags or headers (example 5).



Processing Natural Language in Textual Information Retrieval and related topics.
1. Introduction
"Natural Language Processing" (NLP) as a field has been developing for many years. It was formed in 1960 as a sub-field of Artificial Intelligence and Linguistics, with the aim of studying problems in the automated generation and understanding of natural language.
…

Example 5. Document without headers or tags

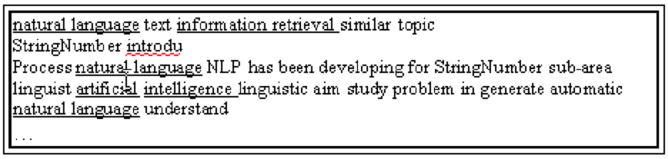
2. Text standardising, consisting in homogenising the whole text in the complete collection of documents to be worked on, including the consideration of capitalised or non-capitalised terms, checking specific parameters like numerals or dates; abbreviations or acronyms, eliminating empty words by applying the lists of functional words (prepositions, articles, etc.) identifying N-Grams, (the example's terms and underlined terms). (Example 6).



processing natural language in textual information retrieval and related topics
StringNumber introduction
Processing natural language NLP has been developing for StringNumber sub-area linguistic artificial intelligence aim study problems in the automated generation and understanding of natural language.
…

Example 6. Standardised document

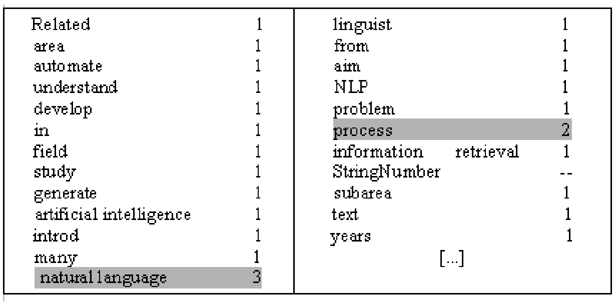
3. Stemming terms is a linguistic process that attempts to determine the base (lemma) of each word in a text. Its aim is to reduce a word to its root, so that the key words in a query or document are represented by their roots instead of the original words. The lemma of word is its basic form along with its inflected forms. For example, "inform" could be the lemma of "information" or "inform." The stemming process (example 7) is carried out by using algorithms that can represent the different variants of a term at once, while also reducing the amount of vocabulary and as a consequence improving the capacity of storage in systems, as well as document processing time. However, these algorithms have the -inconvenience of sometimes not grouping words that should be grouped, and vice versa: erroneously presenting words as equals.



natural language text information retrieval similar topic
StringNumber introdu
Process natural language NLP has been developing for StringNumber sub-area linguist artificial intelligence linguistic aim study problem in generate automatic natural language understand
…

Example 7. Documents with stemmed terms

Parametrising documents consists in assigning a weight to each one of the relevant terms associated to a document. A term's weight is usually calculated as a function of its appearance frequency in the document, indicating the importance of these terms as the document's content description (example 8).



Related	1	linguist	1
area	1	from	1
automate	1	aim	1
understand	1	NLP	1
develop	1	problem	1
in	1	process	2
field	1	information retrieval	1
study	1	StringNumber	--
generate	1	subarea	1
artificial intelligence	1	text	1
introd	1	years	1
many	1	[…]	
natural language	3		

Example 8. Fragment of a parametrised document (see how the frequencies of each term changes as the quantification of the remaining terms in the document continues)

One of the most often used methods to estimate the importance of a term is the TF.IDF system (Term Frequency, Inverse Document Frequency). It is designed to calculate the importance of a term relative to its appearance frequency in a document, but as a function of the total appearance frequency for all of the corpus' documents. That is, the fact that a term appears often in one document is indicative that that term is representative of the content, but only when that term does not appear frequently in all documents. If it appeared frequently in all documents, it would not have any discriminatory value (for example, it would be absurd to represent the content of a document in a recipe database by the frequency of the word food, even though it appears often).

Finally, and as we have already mentioned, we must describe two commonly used techniques in the statistical processing of natural language:

a) Detecting N-Grams: this consists in identifying words that are usually together (compound words, proper nouns, etc.) to be able to process them as a single conceptual unit. This is usually done by estimating the probability of two words that are often together make up a single term (compound). These techniques attempt to identify compound terms such as "accommodation service" or "European Union."

b) Stopwords lists: a list of empty words in a terms list (prepositions, determiners, pronouns, etc.) considered to have little semantic value, and are eliminated when found in document, leaving them out of the terms index to be analysed. Deleting all of these terms avoids document noise problems and saves on resources, since in documents few elements are repeated frequently.

3.2. Linguistic processing of natural language

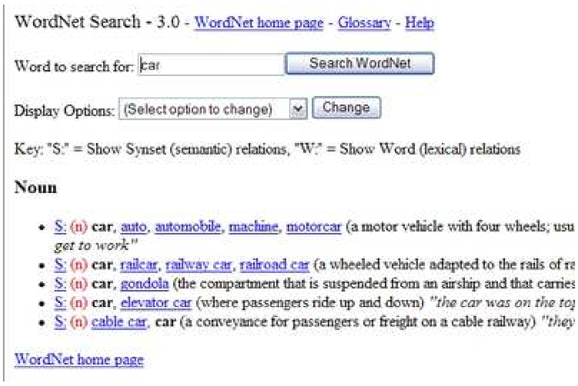
This approach is based on the application of different techniques and rules that explicitly encode linguistic knowledge [Sanderson, 2000]. The documents are analysed through different linguistic levels (as previously mentioned) by linguistic tools that incorporate each level's own annotations to the text. Below we show the different steps to take in a linguistic analysis of documents, even though not all systems use them.

The morphological analysis is performed by taggers that assign each word to a grammatical category according to the morphological characteristics found.

After having identified and analysed the words in a text, the next step is to see how they are related and used together in making larger grammatical units, phrases and sentences. Therefore a syntax analysis of the text is performed. This is when parsers are applied: descriptive formalism that demonstrate the text's syntax structure. The techniques used to apply and create parsers vary and depend on the aim of the syntax analysis. For information retrieval it is often used for a superficial analysis aiming to only identify the most meaningful structures: nominal sentences, verbal and prepositional sentence, values, etc. This level of analysis is usually used to optimise resources and not slow down the system's response.

From the text's syntax structure, the next aim is to obtain the meaning of the sentences within it. The aim is to obtain the sentence's semantic representation from the elements that make it up.

One of the most often used tools in semantic processing is the lexicographic database WordNet. This is an annotated semantic lexicon in different languages made up of synonym groups called synsets which provide short definitions along with the different semantic relationships between synonym groups.



WordNet Search - 3.0 - WordNet home page - Glossary - Help

Word to search for: car [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Noun

- S: (n) car, auto, automobile, machine, motorcar (a motor vehicle with four wheels; usu *get to work"*
- S: (n) car, railcar, railway car, railroad car (a wheeled vehicle adapted to the rails of ra
- S: (n) car, gondola (the compartment that is suspended from an airship and that carries
- S: (n) car, elevator car (where passengers ride up and down) *"the car was on the to*
- S: (n) cable car, car (a conveyance for passengers or freight on a cable railway) *"they*

WordNet home page

Image 2: An example of semantic information provided by WordNet. http://wordnet.princeton.edu/perl/webwn

4. Related topics in research

There are different fields of research relative to information retrieval and natural language processing that focus on the problem from other perspectives, but whose final aim is to facilitate information access.

Information extraction consists in extracting entities, events and existing relationships between elements in a text or group of texts. This is one way of efficiently accessing large documents since it extracts parts of the document shown in its content. The information generated can be used as knowledge and ontology databases.

Summary generators compress a text's most relevant information. The techniques most often used vary according to the rate of compression, the summary's aim, the text's genre and language (or languages) of the original text, among other factors.

Question answering aims to give a specific response to the formulated query. The information needs must be well-defined: dates, places, etc. Here the processing of natural language attempts to identify the type of response to provide (by disambiguating the question, analysing the set restrictions, and the use of information extraction techniques. These systems are considered to be the potential successors to the current information retrieval systems. START natural language system is an example of one of these systems.

Retrieving multi-language information involves the possibility of retrieving information even though the question and/or documents are in different languages. Automatic translators are used on the documents and/or questions, or the use of interlingua mechanisms to interpret documents. These systems are still a great challenge to researches since they combine two key aspects of the Web's current context: retrieving information and processing multilingual information.

Finally, we must cite the automatic text classification techniques, which automatically assign a set of documents into categories within predefined classifications. The correct description of the document's characteristics (usually through the use of statistical techniques -- pre-processed and parametrisation) strongly influences the quality of the grouping/categorization by these techniques.

5. Conclusions

With the aim of understanding the current Natural Language Process, we have concisely defined the key concepts and techniques associated with this field, along with some simple examples to help the reader better understand.

Moreover, we have shown how, despite its years of experience, NLP is a very live and developing field of linguistics, with its many challenges still to overcome due to natural language's ambiguity.

We have paid special attention to the differences between statistical and linguistic methods in natural language processing. Even the scientific communities that support each approach are usually at odds, and NLP is often applied by using a combination of techniques from both approaches. Our experience in this field has made us conclude that it is not possible to claim one approach is better than the other; this even includes the use of a mixed approach.

Relative to information retrieval, the statistical processing techniques are more often used in commercial applications. However, in our opinion, the behaviour and efficacy of the different NLP techniques vary depending on the nature of the task at hand, the type of documents to analyse and the computational cost to assume.

Overall, we can deduce the need to continue working on this with the hope of creating new techniques or focuses that will help us overcome these existing shortcomings. This is the only way we can finally reach what seems like the impossible dream of automatic comprehension of natural language.

Finally, and as an Annexe (Annexe 1), we have described some of the unique aspects of processing Spanish, including the mention of some of the key initiatives developed to process this language.

6. Acknowledgments

This project has been partially financed by the Ministerio de Educación y Ciencia (Spain) as part of the HUM2004-03162/FILO project.

7. References

[Allan, 1995] J. Allan [et al.]. (1995). Recent experiments with INQUERY, in: HARMAN, D.K. from: The Fourth Text Retrieval Conference, NIST SP 500-236, Gaithersburg, Maryland.

[Allan, 2000] Allan, J. (2000). NLP for IR - Natural Language Processing for Information Retrieval http://citeseer.ist.psu.edu/308641.html [26-2-2007].

[Baeza-Yates, 1999] Baeza-Yates, R. and Ribeiro-Neto, Berthier. (1999). Modern information retrieval. Addison-Wesley Longman.

[Baeza-Yates, 2004] Baeza-Yates, R. (2004). Challenges in the Interaction of Information Retrieval and Natural Language Processing. in Proc. 5 th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2004), Seoul , Corea. Lecture Notes in Computer Science vol. 2945, pages 445-456, Springer.

[Carmona, 1998] J. Carmona [et al.]. (1998). An environment for morphosyntactic processing of unrestricted spanish text. In: LREC 98: Procedings of the First International Conference on Language Resources and Evaluation, Granada, España.

[Figuerola, 2000] C. G. Figuerola. (2000). La investigación sobre recuperación de información en español. In: C.Gonzalo García and V. García Yedra, editors, Documentación, Terminología y Traducción, pages 73-82. Síntesis, Madrid.

[Figuerola, 2004] C. G. Figuerola [et al.]. (2004). La recuperación de información en español y la normalización de términos, in: Revista Iberoamericana de Inteligencia Artificial, vol VIII, nº 22, pp. 135-145.

[Manning, 1999] Manning, C. D. and Schütze, H. (1999). Foundations of statistical natural language processing. MIT Press. Cambridge, MA: May, p. 680.

[Rodríguez, 1996] S. Rodríguez y J. Carretero. (1996). A formal approach to Spanish morphology: the COES tools. En: XII Congreso de la SEPLN, Sevilla, pp. 118-126. http://www.datsi.fi.upm.es/~coes/

[Sanderson, 2000] Sanderson, M. (2000). Retrieving with good sense, In: Information Retrieval, 2, 49-69.

[Santana, 1997] O. Santana [et al.]. (1997). Flexionador y lematizador automático de formas verbales, In: Lingüística Española Actual, XIX(2), pp. 229-282. http://protos.dis.ulpgc.es/

[Santana, 1999] O. Santana [et al.]. (1999). Flexionador y lematizador de formas nominales, en: Lingüística Española Actual, XXI(2), pp. 253-297. http://protos.dis.ulpgc.es/

[Strzalkowski, 1999] Strzalkowski, T. (1999). Natural Language Information Retrieval. Netherlands: Kluwer Academic Publishers.

[Vilares, 2005] Vilares Ferro, J. (2005). Aplicaciones del procesamiento del lenguaje natural en la recuperación de información en español. http://coleweb.dc.fi.udc.es/cole/library/ps/Vil2005a.pdf [11-3-07]

8. Annexe 1. Unique cases in Spanish natural language processing

As an annexe we will go on to show some of the most important characteristics of natural language processing in Spanish:

1. Empty words lists [Figuerola, 2000]: creating these tools for Spanish is quite a challenge, mainly due to the lack of collections and statistical studies in Spanish that would advise for or against its use. Furthermore, creating these lists varies in function of whether or not they are used in processing general or specific information. If our corpus is not field specific, then the list of empty words should mainly include: determiners, pronouns, adverbs, prepositions and conjunctions. But if the information to be analysed is field specific, this list should be modified and/or extended by an expert of the field in question. We must also mention that many researches have noted the advantage of using fixed expressions as elements in empty words lists. Specifically [Allan, 1995] recommends using a short list of "empty phrases": indication of, which are, how are, information on.

2. Stemming techniques: the majority of information retrieval techniques use frequency counts of terms found in the documents and queries. This implies the need to standardise these terms in order for the count to be carried out properly, taking into consideration those terms with the same lemma or root. There are various lemma and morphological analysers for Spanish. Among them it is worth highlighting the following: COES [Rodríguez, 1996], open to the public with a GNU license; the morphosyntatic analyser MACO+ [Carmona, 1998], or lemma finders FLAMON [Santana, 1997] / FLAVER [Santana, 1999].

Finally, it is worth noting that experiments of this kind of algorithms in Spanish has shown that standardising terms through stemming techniques provides improved results. The S-stemmer algorithm comes up with surprising results. This algorithm is very simple and basically it simply reduces plural words to their singular form. In its original version (for English), this algorithm only eliminates the last "s" of each word. For Spanish, this algorithm could be reinforced by including plural forms of nouns and adjectives with consonants, thus ending in "es." Eliminating the "es" suffixes could produce inconsistencies with words ending in "e" in their singular form, which would require the elimination of "e" endings. We have also shown that eliminating gender specific "a" or "o" endings improves results.

The greatest advantage of this algorithm is its simplicity. However, the inconvenience is that the S-stemmer is incapable of distinguishing nouns and adjectives from other grammatical categories, thus applying it to all words; it also does not distinguish between irregular plural forms. But on the other hand, by treating all words in the same form, it does not introduce additional noise.

Last updated 15-03-2011

© Universitat Pompeu Fabra, Barcelona