

White Paper Lingway

Natural Language Processing : the key of texts structuration

**Information extraction,
Metadata generation,
and automatic Indexation:**

Table of contents

1	THE INFORMATION EXTRACTION: A NEW NEED	2
2	THE INFORMATION EXTRACTION METHODS	4
2.1	THE UNDERSTANDING PHASE: THE LINGUISTIC ANALYSIS	4
2.2	THE SELECTION PHASE	6
3	THE INFORMATION EXTRACTION IN LINGWAY KM	7
3.1	LOCALIZING THE LOGICAL STRUCTURE OF THE DOCUMENT	7
3.2	EXTRACTING NAMED ENTITIES	7
3.3	EXTRACTING THE THEMATIC DESCRIPTORS	7
3.4	EXTRACTING RELEVANT SENTENCES	8
3.5	EXTRACTING ATTRIBUTES AND FACTS	8
3.6	EXTRACTING MULTILINGUAL TERMINOLOGIES	9
3.7	EXTRACTING ASSOCIATIONS BETWEEN EXTRACTED OBJECTS	9
4	CONCLUSION	11

1 The information extraction: a new need

The need for structuring textual information is as older as the documentary issue: for a long time indexation methods, documentary languages and thesaurus have been the basic tools of information officers. To these approaches of documents' manual indexation, which are slow and expensive and therefore only suitable for reasonably sized databases, were often opposed the "full text" search engines which do not require a thematic indexation and carry out the search directly within the text, methods allowing the processing of important volumes even if it implies a loss of quality.

During these last years, the evolution of great importance has been an increasingly large automation of the indexation phase itself. This change has been done under the pressure of several factors:

- The unceasingly increasing volume of data to be processed obliges to search automatic methods.
- The evolution related to semantic web, which tends to structure more information with the creation of new languages (RDF, OWL) and ontologies which aim at enabling the interoperability between documentary sources.¹
- The increasing demand for tools related to economic intelligence requiring more precise tools than traditional statistical or Boolean search engines in order to better localize interesting information and new associations between information.
- The increasing need for processing multilingual information which requires tools for extraction of correspondences between languages.
- The increasing demand for reading assistance tools, as a consequence of the increasing volume of available textual information: fast navigation within a text, automatic summary, and relevant passages identification are as much functionalities which can be correctly applied only on analyzed texts.

The techniques of information extraction (**"IE" for Information Extraction**), based on technologies of automatic processing of languages (NLP) makes it possible to structure a textual information which is at the beginning devoid of any logical structure.

¹ Please refer to « No semantic Web without an automatic processing of languages », whitepaper available on www.lingway.com

The IE field is often broken down in several sub-problems:

- Extracting named entities
- Extracting thematic descriptors (free or standardised)
- Extracting relevant sentences according to a given point of view
- Extracting attributes
- Extracting associations between named entities and descriptors
- Extracting multilingual correspondences

More specific issues, such as the resolution of co references are often related to the IE. Even if this type of problem does not have a global solution yet, Lingway takes part in research tasks on these subjects.

Finally, although it does not act to be strictly accurate of information extraction, methods of logical structuring of the text – to localize the title, the different parts, etc. – are rather close and refer to the same kind of techniques.

2 The information extraction methods

The extraction of information from a text requires two kinds of processing: an « **understanding** » phase and a « **selection** » phase. Some purely statistical tools do not take into account the understanding phase, which usually leads to debatable results.

Indeed, the traditional methods which do not rely on linguistic techniques tend to find “too many” descriptors, especially because descriptors having the same meaning, which are derived words, are not unified.

In Lingway KM, the extractor of *themes* or *descriptors (documentary objects)* is based on the recognition of *terms (linguistic objects)* which are localized by their own linguistic structure: for instance, a term formed in English on the structure “Noun-preposition-Adjective-Noun?” such as “*raising of wages*” or “*raising of minimum wages*” is a good candidate for being a thematic descriptor. It will be actually retained or not to index the document on the basis of additional criteria.

If the extractor finds two descriptors such as « *action of the government*” and “*governmental action*”, they will be automatically unified under one of the two forms, which contributes obviously to a much more effective indexation than the one retaining the two forms.

This unification can be carried out on the basis of morpho-syntactic criteria, such as in the above-mentioned example (« *governmental* » is an adjective derived from the noun “*government*”), but can also be executed on semantic criteria from data of dictionary (“*action of the White House*” could be unified with “*action of government*”).

2.1 The understanding phase: the linguistic analysis

Progress could be made in the field of NLP from the day onwards the problem, too wide to be approached overall, could have been broken down in several sub-problems corresponding with sub-tasks articulated between them. This decomposition is carried out at the same time on the level of linguistic resources (dictionaries and grammatical rules) and on the level of algorithms (analyzers, transducers, etc.).

The main traditional levels are:

- the morphological level: to identify the words of a sentence.
- the syntactic level: to identify the constituents and the functions of a sentence.
- the semantic level: to identify the meaning of the words and the logical structure of a sentence.
- the text level: to identify the structure of a text and the relations between sentences.
- the corpus level: to identify the structure of a set of texts.

Each task (level) can also be broken down in sub-tasks.

Thus the morphological level can be re-broken down in:

- the tokenisation: to identify the frontiers of words (simple and compound) and of sentences.
- the tagging: to identify the category – noun, verb, adjective – of each word.
- the lemmatisation: to identify the canonical form in the dictionary.

The syntactic level is often broken down in:

- the chunking: to identify the frontiers of great importance of constituents (group of nouns, verbs, etc.) and/or relations of great importance between the words.
- the functional tagging: to assign grammatical functions to constituents.
- the parsing: to build a tree representing the structure of the complete sentence.

The semantic level is also broken down in:

- the selection of meaning (WSD « word sense disambiguation »): to select the meaning of each word.
- the logical structuring: to identify arguments of each predicate and their semantic role (agent, purpose, place, etc.).

The text level can also be broken down in:

- resolving anaphors (antecedents of the pronouns, ellipses, references).
- defining the rhetoric structure (comments, explanations, causalities, etc.).
- defining the thematic structure (what about is the text?)

Finally the corpus level can be broken down in:

- defining the nature of documents (press article, technical article, lawful text, commercial booklet, etc.)
- the thematic structure of the corpus (what about is the corpus?).

Most of these tasks are related to sub-disciplines of the linguistic. According to the NLP engineering point of view, they correspond with modules, with specific programs, each one of them needing a type of linguistic resources and a specific algorithmic type.

On various levels, all these modules are implemented and used in the Lingway's tools.

2.2 The selection phase

In many use cases, the extraction phase is not enough. It is still necessary to sort out what is interesting and what is not, which usually depends on the context of use.

In that phase, we first use statistical indicators on the frequency of words in the global language, in order to give more or less relevance to extracted terms and to only keep the most relevant ones. For instance, in the previous paragraph, the term “use case” (Noun--Noun) is related to the reason given in the beginning of that section. Therefore it will thus be a candidate for being descriptor of the text. But as the noun “case” is very frequent in English, the term will get a weak “score” and could be rejected (in first approach) by the selection phase on behalf of terms carrying more information.

Nevertheless it is a first level of selection which needs to be refined according to the text, and even the corpus. Indeed, in a text of IT specifications the term “use case” will be effectively relevant. In that example, other statistical indicators based on its frequency in the text or on the number of co-occurrences of both words (“case” and “use”) will modify the score and lead to consider “use case” as a good descriptor to be retained for this document.

Finally, the corpus itself has a role of modulation on a superior level, by using similar methods. For instance, if the analyzed corpus is a compilation of all of the “use cases” of a company, then the term becomes irrelevant and will be rejected. If it concerns a wider context, then the term will enable to quickly select a sub-set of documents.

Generally speaking, less than 10% of terms extracted on linguistic basis are retained as relevant descriptors by Lingway KM. Therefore this selection phase is very important in the process. However, in order to let “work” such statistical indicators, it is crucial to apply them to an advanced linguistic process: on one hand a selection can not create “relevant” terms if it is applied on a wrong extraction; on the other hand, statistical measures are very different as we consider or not, for instance, “governmental action” and “action of the government” as separated entities unrelated the one to the other. Where countings on character strings offer a mediocre quality in number of non-linguistic systems, they provide all their power on entities linguistically and semantically based.

It is the combination of advanced statistical algorithms with a powerful linguistic extraction which makes the force of Lingway KM.

3 The information extraction in LINGWAY KM

The following paragraphs describe the types of information extracted by Lingway KM as well as the way they are used.

3.1 *Localizing the logical structure of the document*

In using both the text structure (HTML tags, typography, etc.) and the analyzing rules of textual content, the system can localize:

- sequences such as summary, introduction, conclusion, bibliography, acknowledgements, appendix, etc.
- the presentation of a result, an advantages/drawbacks comparison, a prices comparison, etc.

These modules are usually specific to a given type of document (a scientific article, an administrative circular, a job offer, etc.).

3.2 *Extracting named entities*

The identification of « named entities » has been identified as a separated task during the MUC (Machine Understanding Conference) evaluation campaign conducted ten years ago. Three sub-tasks are usually differentiated: recognition of named entities (organizations, persons, and places), temporal expressions (dates and other temporal markers) and numerical expressions (measurable importance, quantities, percentages, etc.).

Lingway makes it possible to automatically extract this kind of entities:

- persons: essentially on the basis of rules of type “first name + X” or “X chairman of Y”. The extractor of persons’ name identifies the surname, the first name, the position if available.
- places: essentially based on files of places name. The extractor identifies the name of the place, its type (country, town, state, etc.).
- organizations: based on lists and rules. The extractor identifies the name of the organization and its type (public limited company, association, etc.).
- dates and other temporal markers.

3.3 *Extracting the thematic descriptors*

We usually distinguish several kinds of thematic indexation:

- Controlled indexation: only terms belonging to a pre-established standardized list known as “authority list” can be retained. The word of “categorization” is also often used to indicate this kind of processing.
- Free indexation: the terms selected as thematic descriptors can be freely selected independently of any authority list.

The problems related to automatic indexation of a text with thematic descriptors are quite different according to the use of one or the other of these indexation methods. However in both of them is first carried out a phase of identification of the “themes”, then a possible processing for regrouping or normalizing these terms and finally a relevance calculation specifying the significance of the theme in the document.

This normalization can be carried out in the frame of a free or controlled indexation.

3.4 *Extracting relevant sentences*

The relevant sentences are localized by the presence of typical “linguistic markers” – meaning sequences of typical words of a type of sentence – possibly completed by the identification of the part of the text to which belongs the sentence, and finally of relevant words in the context (for instance, a word of the title belonging also to the analyzed sentence).

For instance, relevant sentences are:

- A thematic announcement: meaning a sentence which announces or repeats the themes developed in the document or in a given part. For instance: “***In that text, we will study the implications of our works and nanotechnologies***”, “***Our document is divided into three parts, ...***”.
- An intensification meaning: a sentence which relevance is underlined by the author using rhetoric means. For instance: “*The fullerenes have a **particularly relevant** structure for the medical uses*”, “***I insist on the fact that these techniques do not have any side effects ...***”

A considerable phraseology work has been undertaken to describe these linguistic markers, usually according to a specific type of corpus. The important point is that even if a great number of markers can exist for a given type of sentence, they are in most of the cases in finished number and then can be described.

A methodology has been set up which makes it possible to compile these lists by using various tools for analyzing the corpus as well as general dictionaries.

3.5 *Extracting attributes and facts*

The extraction of entities makes it possible to identify persons, organizations, places, etc. However we usually want to go further and to associate attributes extracted from the text to these entities: the *position* of a person (director, project manager, etc.), its *role* in a document (author, quoted person, person responsible for a file, addressee, etc.) or the *nature* of an organization (company, association, trust, etc.) its *turnover*, etc.

In Lingway KM, contextual rules allow to extract these associations and to generate an appropriate XML/RDF description.

For instance, regarding an application in the employment field, Lingway KM makes it possible to identify within the text of a job offer various elements describing the offer, such as the company’s *size*, the *reason* for the recruiting, the required *experience*, the offered *hierarchical position*, etc.

3.6 *Extracting multilingual terminologies*

Lingway KM makes it possible to build automatically a bilingual terminology, without requiring an aligned corpus, as it is mostly the case in other approaches. This functionality is especially useful when Lingway KM is used jointly with a search engine dedicated to machine translation.

The general principle is the following:

The terms are extracted from the corpus in the source language. Each term is sent like a « cross-language » query to a corpus in a target language, which will have been beforehand made up, generally by crawling the web. This corpus must obviously belong to the same field as the corpus in source language, but does not require to be aligned.

The corresponding terms in the corpus in target language are as many candidates for being translations. That – or those – which most frequently appear are selected like translations of the starting terms.

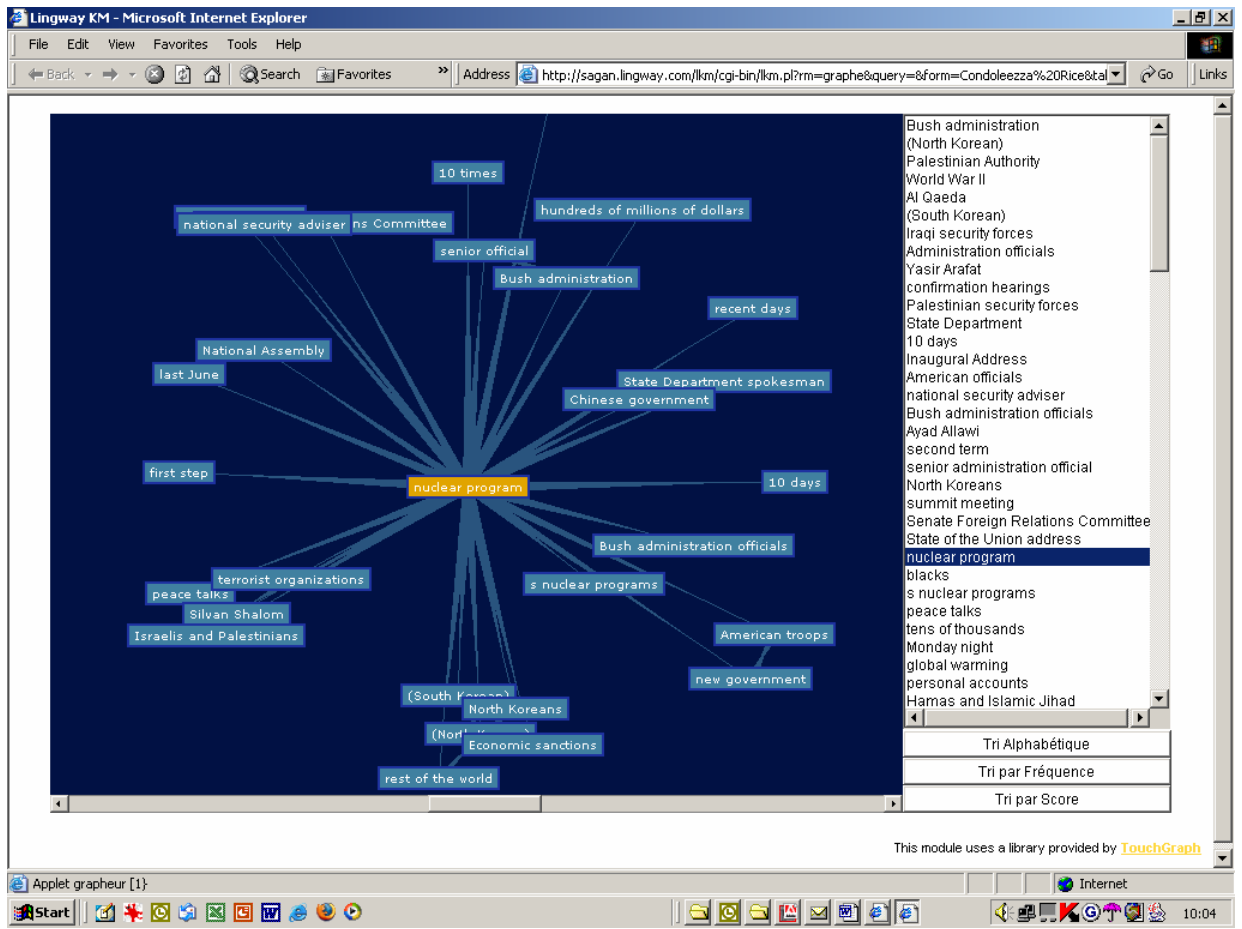
This method makes it possible to very quickly build terminologies of several thousands of terms, which is especially convenient in the following cases:

- Dictionary creation for a machine translation engine.
- Multilingual categorization, offering the possibility to categorize in other languages than the one used during the training.
- Development of the glossary of the company.

3.7 *Extracting associations between extracted objects*

The demand for tools of visualization is increasing. One of the functionalities usually required is the localization of associations between named entities and/or thematic descriptors identified in the texts. So that this type of approach is useful, the extracted objects must necessarily be relevant: some visualization tools carry out associations between simple words which has only little interest.

The graph below shows associations extracted from the press in January-february 2005. A query has been done with “**Condoleeza Rice**”. The following graph shows the semantic environment of different themes, in those documents where answering that query: In this picture, we have focuses on the term “nuclear program”, showing several sub-clusters, around Korea, around Middle East conflict, around China, around US administration. Clicking on one link gives access to the documents from where the association has been discovered.



4 Conclusion

The information automatic extraction, and more widely the issue of structuring and describing textual information, becomes an important functionality of the information systems, because unceasingly increasing volume of information can not be indexed manually any more and because the information mass offered to each reader becomes so important as reading assistance tools are needed.

The methods based on statistical techniques are necessary but non sufficient. To carry out this task satisfactorily, it is necessary to implement methods of linguistic analysis which imply the use of important automations, grammars and dictionaries. The modules of extraction and structuring of Lingway KM are based on this approach.