An Abstract System for Converting and Recovering Texts like Structured Information

Edgardo Samuel Barraza Verdesoto (edglibre@gmail.com)
UNADE/FABA

Richard de Jesus Gil Herrera
UNADE/UNIR

Marly Yaneth Rojas Ortiz
FABA

Research Article

Keywords: structured information, algebraic structure, context, nominal cores, verbal cores

Posted Date: September 16th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-906361/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
An Abstract System for Converting and Recovering Texts like Structured Information

Edgardo Samuel Barraza Verdesoto\textsuperscript{1,2,3*}, Richard de Jesus Gil Herrera\textsuperscript{1,4} and Marly Janeth Rojas Ortiz\textsuperscript{2}

\textsuperscript{1}\textsuperscript{*}Universidad Americana de Europa (UNADE), México.  
\textsuperscript{2}Research Department, Tecnológica Autónoma de Bogotá (FABA), Colombia.  
\textsuperscript{3}Universidad de Santander (UDES), Bogotá, Colombia.  
\textsuperscript{4}Universidad Internacional de la Rioja, Spain.

*Corresponding author(s). E-mail(s): edgardo.barraza@correo.faba.edu.co; direccioninvestigacion@faba.edu.co;

Contributing authors: richard.dejesus@unir.net;

Abstract

This paper introduces an abstract system for converting texts into structured information. The proposed architecture incorporates several strategies based on scientific models of how the brain records and recovers memories, and approaches that convert texts into structured data. The applications of this proposal are vast because, in general, the information that can be expressed like a text way, such as reports, emails, web contents, etc., is considered unstructured and, hence, the repositories based on a SQL do not capable to deal efficiently with this kind of data. The model in which was based on this proposal divides a sentence into clusters of words which in turn are transformed into members of a taxonomy of algebraic structures. The algebraic structures must comply properties of Abelian groups. Methodologically, an incremental prototyping approach has been applied to develop a satisfactory architecture that can be adapted to any language. A special case is studied, this deals with the Spanish language. The developed abstract system is a framework that permits to implements applications that convert unstructured textual information to structured information, this can be useful in contexts such as Natural Language Generation, Data Mining, Dynamically generation of theories, among others.
1 Introduction

Natural Language Processing (NLP) comprises several applied disciplines which have in common the interaction human-machine; some topics such as the treatment of the language in any specific feature such as voice, text, images, etc., all of these have been very studied [1]. The written communication is a type of information that has a basic structure well defined in each language which is useful in information processing [2], additionally, some other types of communication can be totally or partially converted into text whereupon the processing is carried out [3, 4].

According to the Psycholinguistic field, language processing can be analyzed utilizing two dimensions: syntactic processing and perceptual processing. Syntactic processing refers to the process of understanding a particular sentence and perceptual processing refers to the direct acquisition of environmental information that is contained in the stimulus flux. Additionally, and according to [5], there is a kind of memory for storing and remembering facts and events consciously, this memory is normally classified into two subtypes of memory: semantic memory and episodic memory. An event is stored into the episodic memory such as a storyteller, whereas the semantic memory records the same event as part of our overall knowledge (dictionary). All of these theories summarize that the language is recorded in our brain similar to a text well-written [6, 7].

Several applications have been developed following principles or attributes of the text, for instance, to build ontologies or micro-theories [8–11] which are convenient for automation making decision tasks [12]. In process automation, it is also emerging the urgent need of delegating some decisions in machines because the volume or the speed of the produced information reduce the human performance and delay the production in the decision-making processes [13] [14] [15]. The information retrieval and the generation of natural language from homogeneous or heterogeneous sources is an outcome of the aggregation of a large amount of data generated by interactivity, shared repositories, and the continuous requirements in areas such as basic and applied science, as well as in engineering; or, also, by establishing relationships between the different investigations that are carried out continuously by comparing, deciding or mixing of data [16] [17] [18]. The sources of information on the Web can be heterogeneous which adds more complexity in the design of decision-making systems. Languages such as OWL or C-OWL have devised a path towards the standardization of these logical systems [19, 20], but now, they are presenting certain drawbacks because some ontologies could change or coexist with some others that are, apparently contradictory or introduce ambiguity [21], this happens in both cases with unstructured information, and texts. Regarding
the automatic generation of ontologies from a text, several proposals have been presented [22–27], they have tried to solve the difficulties described above, but with partial success. Creating repositories where the information recorded in them does not change the original structure of a text or, at least, has a more suitable structure for effective extractions can be a solution for these difficulties.

An approach that allows to record and recovery fragments or whole texts from a repository can help to generate new sentences as in Natural Language Generation (NLG) [28], for conceiving or improving strategies applied in recovering of unstructured information [29], or to build instantaneous small theories capable to decide regarding the recovered text, e.g., to generate legal ontologies from texts [30].

This article aims to introduce an architecture that grants to build applications capable to dissociate texts/sentences in subsets of cores with properties and simple operations such as those that the algebraic groups incorporate. The operations that incorporate the algebraic groups are preferable because they have properties that promote an easy manner to retrieve a whole text or part of it by keeping the structure of the language. The relationship that can be established between sets will be analyzed because this plays an important role in maintaining the meaning of the recovered text/sentences.

This paper is organized as follows. Firstly, a theoretical framework under which the proposal is based will be explored. Secondly, a general architecture will be proposed that explains each one of the elements exposed in the theoretical framework. Thirdly, an approach based on the architecture presented applied to the Spanish Language will be analyzed. Finally, the findings and future works will be exhibited.

1.1 Conceptual framework

1.2 Object-action dissociation/integration

Historically, the dissociation of the information by the human brain was observed when comparing Broca’s agrammatical patients, whose speech involves the use of very few verbs in contrast with other anomic patients that had great difficulty finding concrete nouns. Initially, the major difficulty with verbs for Broca’s patients was interpreted based on the highest syntactical complexity of verbs compared to nouns [31–33]. However, the idea that verbs are, in general, harder to produce has been undermined in other studies where it is indicated that patients with anomic difficulties produce verbs more easily than nouns [34, 35]. From a Neurophysiological point of view, there are different opinions and theoretical proposals [36], three hypotheses have been put forward regarding verb-noun storage issues within neural networks: partial separation of verbs and nouns [37, 38], word separation based on morphosyntax [39], and separation between actions and objects [40, 41]. Psycholinguistics also agrees that the distinction between various grammatical categories, particularly between verbs and nouns exist in the brain and
An Abstract System for Converting and Recovering Texts like Structured Information proposes three possible starting points or context to access the information: availability of information related to grammatical class, required grammatical knowledge, and independence between the definition of grammatical class and the semantic differences [42–45]. A considerable number of studies have dealt with aspects associated with the dissociation of the information within the human mind and the conclusions are similar [46].

The counterpart of the dissociation process is the integration process. According to [47], grammatical information is relevant to understand and produce sentences, but a plausible conclusion suggests that the grammatical class information is not a lexical property that can be retrieved automatically; instead, this property is likely to play an important role in the context of a sentence. Fundamentally, the role of the grammatical class in sentence processing is modulated by the linguistic differences regarding the way as words of certain grammatical classes are used within sentences. In all languages, verbs commonly require higher processing than nouns at various levels, firstly, because the processing of verbs is about events and could exist many elements that will have to be integrated, secondly, the verb syntax also demands more processing because verbs should be connected to other words to convey their meaning, latter, nouns are linked to single objects, but they might refer to events too and it is necessary disambiguation. In conclusion, the effects of the grammatical class in the retrieval and representation of simple words are more productive when the context is present [48–50].

These studies and approaches suggest that the information holding inside the brain is a set of clusters (cores) that could be affected by the ambiguity and the context in both, the dissociation and integration.

1.3 Structure of the sentences and the text

The word classification has been a normal practice in the linguistic, computer sciences, education, among others (see Figure 1); this practice normally has different targets and results. For example, Bloom’s Taxonomy emerges from projects focused on promote higher forms of thinking in education [51]. Furthermore, as an instance, ConceptNet is a project based on the sense common concept that was conceived as a semantic network containing lots of things that the computers should know about the world [52] [53] [54]. Another example is WordNet which resembles a thesaurus in that words are grouped based on their meanings, the result is a network that can be browseable easily [55] [56] [57] [58].

A text is more complex than simple words, it is a texture that relates, firstly, words to create sentences, secondly paragraphs, and ignoring other structures, finally several paragraphs directly or indirectly (e.g. by means the anaphor) linked between one another result in a text. Each language has rules to build sentences and paragraphs. According to [59], there are various ways to classify and to describe the languages, but a very common is:
• **Subject-Object-Verb (SOV).** This is the most frequent type of word order in spoken languages. Almost all SOV languages make use of postpositions. Most of these languages use adjectives before nouns.

• **Subject-Verb-Object (SVO).** It is the second most common type of word order and has the largest number of speakers worldwide. These languages can be divided into two subtypes: those that use prepositions and those that use postpositions.

• **Verb-Subject-Object (VSO):** These languages use prepositions always; additionally, they represent a relatively small set of languages. Most of these languages use adjectives after nouns.

• **Verb-Object-Subject (VOS):** Only a small number of languages puts subjects after objects. Some of these languages use prepositions and adjectives after nouns, whereas in other cases they use postpositions and adjectives before nouns.

Some approaches use such classifications to divide sentences, expressions, paragraphs, and texts, and producing categories that are used in specific applications [60, 61]. Additionally, other applications use these characteristics in a reverse way, for instance, to build sentences, paragraphs, or concatenate textual expressions from the same or different sources for generating new expressions; this is being applied in Human Machine Interfaces (HMI) [16].

A text not only has nouns and verbs, else other types of words with different purposes, e.g., emphasizing words, which to join small sentences to produce effects like generalization or itemization, etc. These words play an important role to decide how are the relations between words, sentences, and paragraphs. They can be linked to verbs or nouns, e.g. the determinants which comply with the function of generalization or quantification of nouns [62].
1.4 Algebraic concepts

Modern Algebra is a discipline that deals with the properties of the sets and their elements, also, about the operations that can be executed within them. Modern algebra classifies the sets: semigroups, monoids, groups, rings, and fields, all of these are named algebraic structures. These classifications depend on the number and type of properties the set along with its operations comply. A class more important for this approach is the groups, specifically the Abelian groups because they have significant properties that guarantee that by operating elements of a dissociated sentence, the original sentence can be rebuilt.

For illustration, supposing the following sentence: "Fred wants to go to Hong Kong and to visit tourist places", Figure 2 shows a possible dissociation as a set of clusters:

To know what type of algebraic structure belongs this set, the following properties should be verified:

1. Closure: given two clusters, A and B: A+B = C, where C is another cluster that belongs to the set.
2. Commutative: given two clusters, A and B: A+B = B+A.
3. Associative: given three clusters, A, B, and C: (A+B)+C = A+(B+C)
4. Neutral element (identity): There is an element $\lambda$ such that $A+\lambda=\lambda+A=A$
5. Symmetrical element: given a cluster A, there is a cluster B such that; $A+B=B+A=\lambda$

Before the properties will be validated, the set of clusters generated by the dissociation should define its elements and the possible operations among them. Retaking the example of Figure 2, the generated set contains character strings as elements, thus, an operation possible is the concatenation (+). The addition of this type of cluster is not commutative, but it is possible to
prove that complies with the closure and associative properties that define a semigroup which allows generating small sentences such as:

"Fred" + "wants to go to" + "Hong Kong" = "Fred wants to go to Hong Kong"

If a semigroup have the identity element (\(\lambda\)), it is named a monoid, e.g., the identity element for the example is the null string (\(\lambda=\"\"\)):

\[ \text{Fred} + \lambda = \lambda + \text{Fred} = \text{Fred} \]

The commutative property is necessary because the cores could be positioned at the left or right of the operation without change the result, this is possible if the elements are pre-processed and converted to another type of cluster different to simple strings, e.g., a vector of strings, where the order is implicit. Additionally, the operation would be executed by position as follows:

\[ [\lambda, \"wants to go to\"] + [\"Fred\", \lambda] = [\"Fred\", \lambda] + [\lambda, \"wants to go to\"] = [\"Fred\", \"wants to go to\"] \]

To summarize, the dissociation must guarantee that the generated elements comply with the properties of a commutative group (Abelian group), else the recovery of this information could be ineffective. Therefore, the set, its elements, and operations must be chosen adequately.

1.5 Recovery Functions

Based on the sets proposed in the last section and although the groups along with its operation, minimize the production of sentences without sense, some retrieved sentences could have a bad structure such as:

["Fred", "wants to go to"]

To resolve this problem, a recovery system must implement a function that maps this type of expression to [\(\lambda\)], removing them from the final answer. Additionally, the implementation will be different for each language. Therefore, a system multilingual would be similar to a framework that includes functions with this purpose.

2 Methodology

This section proposes an architecture of a system, for dissociating and recovering texts and sentences, based on the concepts, theories, and regulations aforementioned. Figure 3 shows a scheme of the system. The system would include three major sub-systems: dissociation, memory and recovery. The two first sub-systems will be activated serially and immediately a reading takes place, and the latter process is executed when a query promotes the generation of sentences. Nevertheless, in terms of the information processing associated with each sub-system, they operate independently. The entire system is conceptualized as a framework that could be up-gradable and enriched with plug-in modules.
2.1 Dissociation subsystem

The function of the dissociation subsystem is to split a text/sentence into special units. As previously mentioned in 1.3, all languages share a common characteristic that is the identification of three basic clusters within a sentence: Subject(S), Verb(V), and Object(O). They can occur within a sentence in a different order depending on the language.

In this paper, the expression SOV-trio or simply SOV will be used to represent the trio that models a sentence or a text. Given that Subject and Object have similarities both will be treated as (S). Additionally, each of the components of a SOV will be named a core.

The cores may contain one or more words from the sentence. For example, it is possible to have a verb followed by another verb in the same core, as in the following sentence: Fred wants to go to Hong Kong; thus, the two verbs (“to want and to go”) constitute a V core. Once a SOV is generated, this is dispatched to the memory subsystem.

2.1.1 Strategies to generate SOVs

As explained in section 1.2, there is a consensus about the dissociation between actions (verbs) and objects (nouns) inside the human mind. However, [47] emphasizes the existence of drawbacks, for instance, problems by establishing the grammar category that can generate confusion between verbs and nouns, this also can happen in the process of dissociation in this subsystem. To dissociate the sentences correctly, the subsystem should implement modules such as:

```java
// Diagram
```
• **Syntactic Analysis (Parsing).** An ordinary parser generates a syntax tree from which the SOVs can be rapidly built. Although this strategy is good, it does not avoid that the syntax tree generated may require the involvement of some other heuristic processes to “refine” the creation of the cores, for instance, in cases of slang interpretation.

• **Grouping of elements.** The dissociation in cores requires identifying elements like *determinants, adverbs, prepositions, conjunctions*, etc., in such a way that they will be inserted in the adequate core. This process should be customized for each language.

• **Dictionaries and conjugators.** Sometimes, parsers can produce an incorrect word classification, especially when the parser has not well-trained in a particular language, in such case it is necessary to perform an analysis and debugging process over these words. For this purpose, software like dictionaries and conjugators modules could be useful to validate the category.

To summarize, some procedures, syntactic-semantic strategies, and heuristics should be implemented to help in building the S/O/V cores correctly.

### 2.2 Memory subsystem

An important function of the memory system is to store the information generated by the dissociation subsystem; hence, it is mandatory to build a structure that guarantees order and efficiency. Therefore, the memory system should contain a repository to save the SOVs generated by each text read interrelated between them. This storage should maintain these cores in such a way that they could be retrieval in the exact order as they were read. According to these principles, the implementation should comply with the following conditions:

• **Set of S/O/V cores.** The database engine must obey different rules to the traditional linguistic because their detail level is different as was explained in the above section. They should be equipped with algebraic operations (see subsection 1.4) that part or whole original text will be rebuilt simply.

• **Repository based on queries.** A repository based on query means that the store uses SQL technology to save and recover its cores because the SQL systems are efficient, regular, and easy to implement. The queries can be attended by modules based on context to recover more exactly the cores required to create sentences, paragraphs, and full texts. Furthermore, the modules must guarantee that the retrieved information has a high degree of compatibility with the query. This latter is possible if the information is organized in a structure like Wordnet, but rearranged to manage S/V/O cores (and no words).
An Abstract System for Converting and Recovering Texts like Structured Information

2.2.1 Dynamic structure

The sequence of SOVs can be re-organized in distributed sets connected in terms of their original semantic content. Figure 4 shows a scheme that could be used for this purpose; this will be a digraph where each node contains a set of SOVs (Abelian group).

The nodes will be related with adequate functions (mapping) to guarantee that the recovery of the part, or the whole, of a sentence/text will be executed correctly.

2.3 Recovery subsystem

The purpose of this subsystem is to generate, in a dynamic way, a sentence/text part or entirely. This subsystem is closely interrelated to the dynamic structure because this subsystem is composed of the functions that connect the nodes.

2.3.1 The Engine

The queries are expected in natural language and it would transform into a set of SOVs. The idea is to compare SOVs for finding the most close results.

The strategies to match the SOVs can be wide. An example could be to establish matches of SOVs that contain elements that could respond contextually to the query as in Figure 5. The degree of coincidence will be the measure.

This strategy could recover sentences that do not answer the query completely, hence, it would be important to implement another stage, for instance, that compares sentences in a logical context. This can be carried out by converting the query and the text recovered into small theories that can be matched logically.

3 An approach

An approach for the Spanish language of this abstract system has been developed by Barraza et al. [63, 64], this follows, approximately, the directives
proposed previously. This section has a summary of this system and its prototype designed as a **layered framework** that could be used for any language by implementing the respective modules. The most relevant layers of the dissociation processes are the following:

1. **Identifying the language.** This first layer has been designed to identify the language of the text and to divide it into sentences, and finally, their results will send to the next layer one at a time.

2. **Planning.** This second layer chooses the modules required to dissociate the sentences based on the language recognized. This layer makes flexible the framework because it allows changing the rules of dissociation depending on the language to process.

3. **Reaction layer.** This layer executes the modules chosen by the planning layer, it has been named this way because the modules are executed is such as a chain reaction, i.e., the modules could be organized in several pipes and executed simultaneously, obtaining more than one only answer.

The modules created in the reaction layer for the Spanish language implementation have been organized in three linear phases. In the first phase, each sentence is processed by a linguistic tool, commanded by the VISL parser reaction layer. The second phase receives from the parser output a very extensive word classification which is reduced in **SOVs**. Finally, in the third phase, the refined cores are created and saved in a normal database. Each sentence is organized as a hierarchy of Abelian groups with a binary operation capable of rebuilding part or totally the expressions embedded inside them. The Abelian
An Abstract System for Converting and Recovering Texts like Structured Information

Fig. 6 Example of a dissociation process carried out by the approach in [63, 64]

Groups contain, as elements, vectors of couples of vectors of nominal and verbal cores. A vector could not contain verbal cores but always must contain nominal cores (see Fig 6).

Table 1 shows the classification established heuristically for the Spanish language cores in this approach. This process involves a loop where neighboring words that comply with certain conditions are packed into a single class named: **nominal core (S)**, **determinant**, and **verbal core (V)**. A determinant is used to interrelate the Abelian groups as in Figure 4 of section 2.2.1.

Table 1 Final categorization scheme

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p, J$</td>
<td>Determinants</td>
</tr>
<tr>
<td>$v, V$</td>
<td>Verbal core</td>
</tr>
<tr>
<td>$n$</td>
<td>Noun core</td>
</tr>
</tbody>
</table>

In this approach, there are two types of determinants $p$-$det$ and a $J$-$det$; both interrelate the sets with functions, but the Abelian group pointed by a $J$-$det$ is considered optional in the rebuild of the sentence. All of these properties were established in a empirical way.
The restoring process is not the reverse operation exactly, else it is a complex process that execute tasks from the repository trying to preserve syntax and the original semantics. This purpose is successful due the properties of the Abelian groups (see [63]) and the hierarchy of sets created by the determinants in the dissociation process. The process is shown in Fig 7.

The sets are operated and mapped in a domino way from the core where the matching occurs up until the root of the hierarchy. For example, in Fig 7, if the core matched is in $G_3$ and corresponds to $n_2 = \text{críticas}$ then the recovered sentence will be: “La marcha programada, para el próximo 26 de marzo, recibía críticas”.

4 Conclusions and future works

The high consumption demand of information has caused that the automation of processes such as decision-making, pattern recognition, interaction human-machine, among others, becomes very important. Several of these processes require the use of the text, either to understand queries, generate reports, or answer in natural language, hence, to build applications with these functions takes a greater relevance. This paper presents an architecture for dissociating
An Abstract System for Converting and Recovering Texts like Structured Information

the text/sentences, saving it in a SQL database, and recovering it without loss of meaning. This is highly productive in process automation because the textual information is converted from unstructured to structured and the queries and other processes in natural language can be more efficient.

The abstract system has been inspired and based on processes verified by scientists related to the dissociation of the information inside of the human brain, memory models in Neuroscience, and the structure of the languages in such as Linguistic and Psycholinguistic disciplines. The framework proposed divides a sentence/text into clusters like the brain dissociates the speech into nominal and verbal categories. The scheme will divide the text/sentence into sets of cores named nominal cores and verbal cores, and implement an algebraic operation that can be used to generate new sentences without losing the structure of the language. This proposal was applied by the approach studied in the last section successfully.

The explored implementation resolved a great part of the challenges described in the paper by implementing a framework with abstract modules that can be custom implemented, for instance, the processes described in the architecture, the generic abstract modules for different languages, and the recovery modules, among others.

Additionally, the implementation creates a solution for the Spanish language by using heuristics for both dissociation and recovery. The application suggests interrelating the algebraic sets employing functions with the aim to recovery whole or part of the textual information by maintaining the meaning. The approach shows that for the Spanish language is possible to have an implementation.

An abstract system has been proposed for converting unstructured textual information to structured information. This proposal has been tested in an approach for the Spanish language successfully. The future works will be addressed to implement this framework for others languages and to generate applications to these approaches.

References


An Abstract System for Converting and Recovering Texts like Structured Information


An Abstract System for Converting and Recovering Texts like Structured Information


An Abstract System for Converting and Recovering Texts like Structured Information


An Abstract System for Converting and Recovering Texts like Structured Information

1016/0028-3932(94)00066-2


An Abstract System for Converting and Recovering Texts like Structured Information


**Figure 1**

A simple example to match
Figure 2

Recovering a sentences. Extracted from [63]
Figure 3

Example of a dissociation process carried out by the approach in [63, 64]
Figure 4

Directed graph of the sentence: "Fred wants to go to Hong Kong and to visit tourist places"
Figure 5

Some word classifications techniques [51][53][58]

Figure 6
Architecture of the system

C: Set of clusters

Fred \( \text{N} \)

wants to go to \( \text{V} \)

and to visit \( \text{V} \)

tourist places \( \text{N} \)

Hong Kong \( \text{N} \)

Figure 7

Dissociation of a sentence