







# Concept Identification from Single-Documents

José Luis Ochoa-Hernández<sup>(✉)</sup> , Mario Barcelo-Valenzuela ,  
Gerardo Sanchez-Smitz , and Raquel Torres-Peralta 

Department of Industrial Engineering, Universidad de Sonora,  
Blvd. Rosales y Transversal, 83000 Hermosillo, Sonora, Mexico  
{joseluis.ochoa, mbarcelo, gsanchez,  
rtorres}@industrial.uson.mx

**Abstract.** This article presents a method that extracts relevant concepts automatically, consisting of one or several words, whose main contribution is that it does so from a single document of any domain, regardless of its length; however, documents of short length are used (which are the most frequent to obtain on the web) to perform the work. This research was conducted for documents written in Spanish and was tested in multiple randomized domains to compare their results. For this, an algorithm was used to automatically identify syntactic patterns in the document. This work uses the previous work of [1] to obtain its results. This algorithm is based on statistical approximations and on the length of the identifiable patterns contained in the document, applies certain heuristic that can enhance or decrease the patterns' choice according to the selection of one of the 5 methods that are processed (M1 to M5), with these patterns the candidate concepts are obtained, which go through another evaluation process that will obtain the final concepts. This proposal presents at least four advantages: (1) It is multi-domain, (2) It is independent of the text length, (3) It can work with one or more documents and (4) It allows the discarding of garbage or undesirable patterns from the beginning. The method was implemented in 11 different domains and its results range varies between 58%–70% of precision and 25%–46% of recall.

**Keywords:** Concept extraction · Syntactic patterns · Text analysis  
Single-documents

## 1 Introduction

The concepts extraction, despite the fact that it emerged in 1954 based on Harris' distributional hypothesis [2] is an activity that still presents important advances. This activity is applied to a large number of tasks that are being used daily, especially in those tasks that use natural language as a source of information for their development, some very common examples of these are: translations of texts, phrases or words from one language to another as [3] which translates from English to Arabic for sentiment analysis purposes, or [4] which translates from Polish to English with human judgment; the detection of plagiarism in articles, research or books [5–7] which presents a work using syntactic-semantic based Natural Language Processing (NLP) techniques for Unmasking text plagiarism; generation of summaries from large corpus of concrete

domain such as Biomedicine [8] or for single-documents of any domain [9]; in fact, there are a great variety of methods as exposed [10] in this review, as well as the creation of domain glossaries in very common languages such as English where [11] create one based on the text acronyms or in less frequent languages such as Islam [12] who creates an unsupervised concept hierarchy.

Another field in which concept extraction has been applied fairly recently is for the creation, extension or integration of ontologies in different domains. As is the case of [13–15]. Transforming the common web into semantic web [16, 17] is another application because with the current web, too much time is lost in knowing which pages are relevant and which are not [10, 18].

In the most recent area where concept extraction is applied is in decision support tools for doctors in different domains, as is the case of [19] who integrate NLP and Machine Learning Algorithms to categorize Oncologic Response in Radiology Reports or [20] which generate structured reports for cancer from free-text (non-structured) pathology reports.

The concept extraction is therefore a process that continues in development. Both in the latest research and in the initial ones, these are mainly based on having a large number of texts from the same domain in order to identify the principal concepts in a “more precise” or “more efficient” way, either by a model, for other or by combining these as in [19].

The initial problem that arises in almost all investigations is to form a sufficiently robust corpus [21] to be able to apply the different algorithms that exist today to extract the concepts, some examples of this are the NLP techniques that use lexical and syntactic analysis [22], lexical and syntactic patterns [23], the C value/NC Value of [24–26].

## 1.1 Related Works and Distribution

To know if these extracted concepts are relevant, is necessary to use metrics to measure the importance of the concept usually in the CORPUS. One of the most common methods used in this type is the Term Frequency – Inverse Document Frequency (TF-IDF) proposed by [27] and it is used in works like [28]. There are others like Domain Relevance and Domain Consensus, which are measures to know how much a concept is used in a domain CORPUS and to measure the distribution of use from a term in a domain; co-occurrence and others more mentioned in [29] also require a large number of texts to work and provide better results, without generating too much “garbage”. However, in the case of a single document these metrics do not work by their own nature, the metrics of Precision, Recall and F-score are those that can be adapted as commented in [30].

There are multiple studies that refer to work with single-documents. Here we mention three similar works, including one that would be great to try with the single-document methodology. It’s shown because he says he worked hard to get his corpus from a lot of places and the research is very interesting, however this work could have been avoided using this proposal.

The mentioned work is from [31] who presents a work for the extraction of concepts, they discover concepts based on a CORPUS that was “extracted from many

places”, his proposal includes semantic extraction, a statistical filtering and a semantic analysis. It uses manual elements in several parts of its proposal that make it dependent on an expert. The main difference with our work is the CORPUS and the methods that it applies, since this research is based on linguistic patterns and statistical analysis to a single-document.

The majority of the investigations related to single documents are those that work with text summaries, there are very few that only extract concepts, the work of [30] that identifies key concepts using a method similar to ours was found, instead of applying the identification of key patterns, use the concept of k-core on the graph-of-words representation of text for single-document keyword extraction, retaining only the nodes from the main core as representative terms. This approach that takes better into account proximity between keywords and variability in the number of extracted keywords through the selection of more cohesive subsets of vertices. Within the process of Graph-of-words perform a pre-process similar to ours: (1) tokenization; (2) part-of-speech, annotation and selection; (3) stop-words removal; and (4) stemming. And for validate the results they use the metrics of precision and recall with very good values.

In [32] it is proposed a novel approach for extraction of key phrases from single-document without usage of external information. To achieve this extraction, it is necessary to identify a list of hierarchical key concepts (activity of our interest), his analysis is based on the method contained in the area of Formal Concept Analysis (FCA), known as concept lattice, where combination of sentences or paragraphs are used as objects and the presence of terms are attributes. This form hierarchically organized groups of sentences that share the same or similar set of attributes/terms, evaluate their results with various metrics like Term Frequency, TF-IDF, Chi-Square and Information Gain function. Quality of methods is compared with precision and recall metrics.

[33] is another investigation that requires identifying key concepts to create a summary from a single document, however, to achieve it, its methodology is divided into 4 phases, the one that is of our interest is “text pre-processing”, which is divided into 5 stages: segmentation, tokenization, elimination of stop-words, stemming, and building document words stems list. The difference in this stage is that they use any punctuation mark to divide into sentences the text and we do not, they also have a limited stop-word list, which we eliminate with syntactic patterns, therefore, a predefined list will never be required, finally, they generate a list of keywords that do not specify how they got it, which use to evaluate the extracted sentences and then measure them by an adaptation of the TF-IDF metrics in vectors similarly called IDF, TF and TF-IDF, which count the times a term occurs in a sentence.

So “being able to identify concepts from a single document” was what motivated the development of this research. The work was done and shows that the result is promising, since they manage to identify between 58%–72% of the concepts with precision and a 25%–47% of recall, low percentage because there are big differences when working with a single document vs multiple documents [30].

The article is structured as follows, Sect. 2 briefly discusses the tools used to understand this research, Sect. 3 presents the methodology followed in this research and Sect. 4 presents a practical case with 11 different domains, Sect. 5 presents the results obtained and in the last section the conclusions are shown.

## 2 Tools Used in This Investigation

In this research, two tools were used and modified during the development of the project: (1) Freeling for the process of language analysis (unmodified) and (2) Pattern learning, which will obtain a first approximation of the relevant morpho-syntactic patterns and thus obtain the domain concepts (modified).

### 2.1 Freeling

The Freeling tool is a library that provides language analysis functionalities (morphological analysis, named entity detection, PoS-tagging, parsing, Word Sense Disambiguation, Semantic Role Labelling, etc.) for a variety of languages (English, Spanish, Portuguese, Italian, among others) [34].

This tool uses the labels proposed by the EAGLES group to achieve the morpho-syntactic annotation of lexicons (linguistic analysis). This annotation provides a lot of information, an example can be seen in Table 1, which shows the category Adjectives, Adverbs, Articles, Determinants, Nouns, Verbs, Pronouns and some more, can be seen in full on the Eagles Tags website<sup>1</sup>.

**Table 1.** Attributes and values of Eagles tags.

Position	Attribute	Value
1	Category	Adjective
2	Type	Qualifying
3	Grade	Appreciative
4	Gender	Male Female Common
5	Num	Singular Plural Invariable
6	Case	-
7	Function	Participle

Table 1 contains in the first column a position which is used to identify the meaning of the letter on the label, the attribute is used to know what type of information is being read and the value provides the meaning that this attribute has.

### 2.2 Automatic Pattern Learning

Automatic pattern learning is a tool proposed by [1] which shows a set of multi-word morpho-syntactic patterns suggested as a result of the textual analysis that is made to a corpus of any domain, based on a set of guiding patterns to do the work.

The article comments that the only thing that needs to be specified is the length and specialization of the pattern (understanding as length, the size of words that will contain

<sup>1</sup> <http://www.lsi.upc.es/~nlp/tools/parole-sp.html>.

a term to be considered in the selection of candidates). For example, a term of length 3 would be “diseñadores de moda”/“fashion designers”, one of length 2 would be “estrella michelin”/“Michelin Star”. And specialization, refers to the use of Eagles tags. So a specialization degree 2 would be something like “NC – Common Noun” o “VM – Main verb”, a specialization degree 3 would be “NCM – Common Male Noun”, “NCF – Common Female Noun”, etc. and a specialization degree 1 would be “N – Noun”, “A – Adjective”, “R – Adverb”, etc.

So to obtain a list of patterns like those presented in Table 4, it is necessary to specify these two elements in the tool, for example: we want terms of length 2 with grade 2, the pattern guide would be XX·XX, terms of length 3, grade 2, the pattern guide would be XX·XX·XX, activity that in [21] was considered difficult to recognize and extract (multi-word terms).

The code was adapted in order to test how the tool works with a single document instead of a large corpus. In the same way, the code was adapted in a way that it could identify single length patterns, and these could be processed in a better way, since the concepts of length 1 represent the majority.

Five methods are also proposed in [1] (M1, M2, M3, M4 and M5) which obtain different patterns according to their procedure, in this case the methods M1 to M5 were not modified, because they were studied and the decision was made to leave them without any alteration.

### 2.3 Evaluation Method

To evaluate this research we used a typical method that is used in information retrieval, commonly called Precision and Recall, some jobs that have implemented these evaluation forms are [30, 35]. Similarly, F-score was used as an integrating measure, which is considered as a harmonic mean that combines the values of precision and recall [36], his equations are the following:

$$\text{Precision} = \frac{\textit{terms correctly identified by the tool}}{\textit{correct terms of the expert}} \quad (1)$$

$$\text{Recall} = \frac{\textit{terms correctly identified by the tool}}{\textit{candidate terms of the tool}} \quad (2)$$

$$\text{F - Score} = 2 * \frac{\textit{precision} * \textit{recall}}{\textit{precision} + \textit{recall}} \quad (3)$$

It is necessary to clarify that these equations were adapted to be able to compare the achievement level of concept identification. It was done because these metrics are recognized and are easily comparable, in addition to the other metrics such as those mentioned in [29] are applicable to multiple-Documents. In the work done by [33] something similar happened, the TF-IDF metrics were adapted to be able to measure their results.

### 3 Concept Extraction Process from a Single Document

#### 3.1 The Proposal

In this section we explain the elements of the proposal that is made to extract concepts from a single document, it is formed by 4 main components: the Text, the morphological labeling, the pattern learning and the concepts selection. In Fig. 1 it is possible to see this proposal.

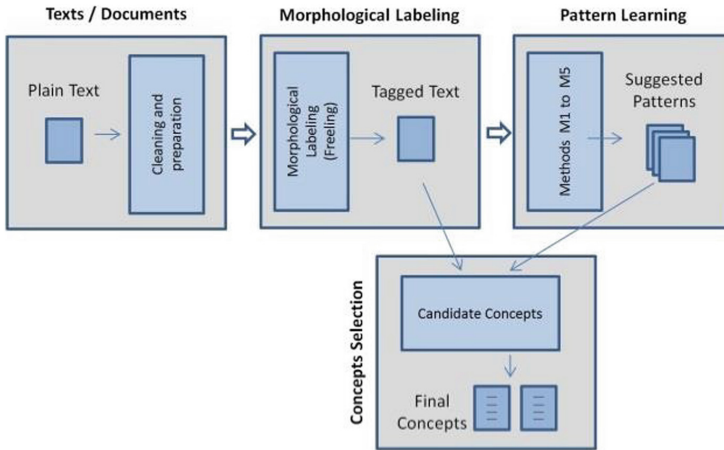


Fig. 1. Proposal of concept extraction process from a single document.

#### 3.2 The Text/Documents

Free text is the source of most research involving NLP. For this purpose, any text of the domain of interest is selected, normally constituted by documents, which can be short in length [32]. The document is transformed into plain text and pre-processed to “*clean the text*” of characters that may cause errors or bad results in the following stages. This is a process that is done manually, in this case, Freeling for example, does not process certain *single or double quotes* which cause errors when tagged. For this, the quotes are replaced by the same ones using another type of character coding, the dashes or long dashes are other elements to review. Likewise, punctuation marks are verified so that sentences, quantities, dates, etc., are well identified.

#### 3.3 Morphological Labeling

Morphological labeling is a key process for this research, Freeling tool is used which performs a text language analysis; it is proven that it does a good job in the labeling process [37, 38]. The result of this process is a list of words morphologically labeled, which during this testing process are reviewed in detail to verify that they have been tagged without errors, in case any word has been mislabeled, given the circumstances

discussed in Sect. 3.1. The source text is modified and re-labeled (see Table 2). In this table it can be seen that the gray marked text, is the text that presented an error when it was labeled.

**Table 2.** Common mistakes in morphological labeling

Text	Labeled	Error type	Solution
'Social Media': menos gurús, más formaciónSilvia Sivera	'·:Fe Social_Media·social_media·NP00000 '·:Fe :::Fd menos·menos·RG gurús·gurús·NP00000 ,·:Fc más·más·RG formaciónSilvia·formaciónsilvia·AQ0FS0 Sivera·sivera·NP00000	Text without separation	Separate words manually
... la intención es que “leáis” con interés ...	la·el·DA0FS0 intención·intención·NCFS000 es·ser·VSIP3S0 que·que·CS ?leái·?leái·VMN0000 con·con·SPS00 interés·interés·NCMS000	Double quotes in “leáis”	Change the quotes for them in another encoding
En septiembre nos “tomaremos” en Barcelona...	En·en·SPS00 septiembre·[??:??/??:??:??:??]-W nos·yo·PP1CP000 ?tomaremo·?tomaremo·VMIC3S0 en·en·SPS00 Barcelona·barcelona·NP00000	Double quotes in “tomaremos”	Change the quotes for them in another encoding

### 3.4 Identification of Key Patterns

Based on the proposed methods (M1 to M5) of [1], this research developed a complete study to know which method provided the best results. The original patterns were obtained and compared with the patterns recommended by the tool, these patterns were analyzed and the method that provided the best results in all cases was the M1 method.

The automatic patterns learning was modified to eliminate those patterns that were useless or unproductive from the beginning, a series of elements that help to configure this method were defined, which improves the accuracy of the results, an example of these are the following (see Table 3):

**Table 3.** Example of unseen patterns of different lengths with an example of concepts.

Length 1		Length 2		Length 3		Length 4	
DI	Un	SP DA	De las	DI AQ NC	Una nueva manera	DI AQ NC SP	
CC	Y	PR DA	Donde el	P0 VM VM	Se está haciendo	P0 VM VM DI	
RG	más	CC NC	Y restaurantes	SP VS AQ	De ser fijo	PR DA NC	
		SP VM	De comprar	PR DA NC	Donde el espacio	VM	
SP	De	DI AQ	Una única	CC AQ SP	Y temporal para	SP DA NC PR	
P0	Se	RG RG	Casi siempre	RG DI SP	Solo algunas de	DA PR P0 VA	
DA	Las					CS DA NC SP	

### 3.5 Selection of Candidate Concepts

With the chosen M1 method (which may vary for other documents) and applied to each document, the results are presented in a list (see Table 4), this list is organized by grammatical categories i.e.: names, verbs, adjectives, etc. in which all the patterns found and suggested by the tool are included.

The tool uses these patterns to identify the candidate concepts of the text. Under normal circumstances of large corpora, it is possible to work with different proposals such as [21]. However, in this new context, identifying the elements in a very precise way is more complicated, the statistical approximation and the heuristic of methods M1 to M5 are the key elements that define the candidate patterns of each document.

### 3.6 Final Concepts Selection

The process of selecting the final concepts includes a cut level according to the method proposed (M1 to M5), which is the minimum level to accept concepts according to their process and the number of times they have appeared in the text, to the list of concepts obtained based on the suggested patterns, the cut level is applied and a final list of concepts is obtained (see Table 4).

**Table 4.** Example of patterns contemplated with their concepts in a document.

Length 1 (AQ)	Length 1 (NC)	Length 1 (NP)	Length 2 (NC AQ)
8·nuevo	7·espacio	6·pop-up	2·casa particular
2·fijo	5·tienda	2·nueva_york	1·experiencia irrepitable
2·itinerante	4·restaurante	2·londres	1·prestigio internacional
2·sorprendente	4·marca	2·nike	1·entorno sorprendente
2·particular	3·concepto	2·barcelona	1·vía alternativo
2·social	3·chefs	1·elisenda_estanyol	1·coste fijo
.....	.....	.....	.....

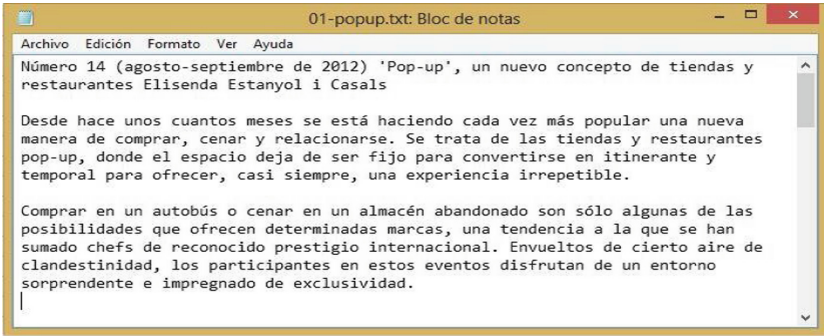
## 4 Case Study - Concept Extraction for 11 Different Domains from a Single Document

This study took as reference the article published by [39] in the journal “Studies of Information and Communication Sciences of the Open University of Catalonia (UOC), COMeIN”, which is entitled “Dissection contents from concepts”. In the article, key concepts are extracted to draw a knowledge map and several examples of different thematic areas/domains are shown, for example: restaurants, online courses, advertising, nostalgia, etc. These domains were randomly selected by the author and they had already identified some concepts and they are related in some way.



### 4.1 The Text/Documents

For this case the 11 articles that are handled in the article [39], were converted to plain text. They were manually pre-processed, eliminating those characters that could cause errors and were left ready to be labeled. An example of this text is shown in Fig. 2.



**Fig. 2.** Plain text of the article – ‘Pop-up’, a new concept of shops and restaurants (<http://comein.uoc.edu/divulgacio/comein/es/numero14/articles/Article-Elisenda-Estanyol.html>).

### 4.2 Morphological Labeling

The Freeling tool was applied to the texts and the result obtained are labeled and lemmatized text, this is shown in the “Labeling” column. For example, given the line “tiendas-tienda-NCFP000” the word “tiendas” is the original text, “tienda” is the lemmatized text and “NCFP000” is the label. A more extensive example can be seen in Table 5, where Text 1 is part of document 01- Creativity and Text 2, is part of document 02-Social media:

**Table 5.** Table with some texts labeled.

Text 1	Labeling	Text 2	Labeling
'Pop-up', un nuevo concepto de tiendas....	'·:·Fe Pop-up·pop-up·NP00000 '·:·Fe ,·:·Fc un·uno·DI0MS0 nuevo·nuevo·AQ0MS0 concepto·concepto·NCMS000 de·de·SPS00 tiendas·tienda·NCFP000	Hubo un tiempo en que para aprender acudíamos a clases...	Hubo·haber·VAIS3S0 un·uno·DI0MS0 tiempo·tiempo·NCMS000 en·en·SPS00 que·que·CS para·para·SPS00 aprender·aprender·VMN0000 acudíamos·acudir·VMN1P0 a·a·SPS00 clases·clase·NCFP000

Table 5 show how the labels provide great value to each word, which facilitates the “creation” of knowledge. For example, in Table 6 we see the scope and specialization of labels:

**Table 6.** Meaning of some labels.

Term	Labeling	Description of the label
Tiendas	NCFP000	It is a Name, Common, in Feminine, in Plural.
Hubo	VAIS3S0	It is a Verb, Auxiliary, Indicative, in the past, in the third person and in the singular.
Un	DI0MS0	It is an undefined determinant, masculine and singular.

### 4.3 Identifying Key Patterns

With the labeled texts, the chosen pattern learning method M1 is applied, to find the patterns. A complete example for the document 01-Creativity of the results obtained can be seen in Table 8, the rest of the suggested patterns for each document are shown in Table 7, omitting the non-suggested ones. It is appreciated that the amount of patterns found in single-documents is smaller compared to large corpora; this is precisely due for the documents size.

**Table 7.** Suggested patterns for the 11 documents analyzed.

02- Social Media	03-Media Planning	04-Advertising	05-Journalism	06-Political Communication
NC	AQ	AQ	NC	NP
NP	NC	NC	NP	NC AQ
NC AQ	NP	NP		NC SP NC
NC SP NC	NC AQ	NC AQ		
	NC SP NC	NC SP NC		
		NP NP NP		
07-Digital Communication	08-Crisis Communication	09-Television	10-Cinema	11-Audiovisual design
NC	AQ	AQ	NC	NC
NP	NC	NC	NP	NP
	NP	NP	NC AQ	NC AQ
	NC AQ	NC AQ	NC SP NC	NC SP NC
	NC SP NC	NC SP NC		

### 4.4 Candidate Concepts Selection

Once the best method has been identified, the patterns found in it are applied separately to each of the texts to obtain the candidate concepts. The result is a list of concepts. At this point, it was proposed to show only the concepts that were best positioned in the study document. However, these concepts were not sufficient to identify all the valid concepts of the document; instead, they were used to identify the most representative concepts of the domain, which are used in other studies as seed concepts [22] (see Table 9).

**Table 8.** Example of suggested and not suggested patterns of the 01-Creativity document.

Method	Suggested patterns		Non Suggested patterns		
M1	AQ·50·*		AO·1	NC·NP·14·*	NC·SP·NC·14·*
	NC·135·*		AQ·NC·22·*	NC·NC·13·*	NC·SP·NP·7·*
	NP·42·*		AQ·NP·5	NP·NC·5·*	NC·NC·NC·5
	NC·AQ·29·*		AO·NC·1	NP·NP·3·*	NC·AQ·NC·4
			AQ·SP·NP·2	NP·AQ·1	NP·DI·AQ·2
		AQ·DI·AQ·1		NP·NC·NP·2	
M2	NC·135·*		AQ·50·*	NP·42·*	NC·AQ·NC·4
	NC·SP·NC·14·*		AO·1	NC·AQ·29·*	NP·DI·AQ·2
			AQ·NC·22·*	NC·NP·14·*	NP·NC·NP·2
			AQ·NP·5	NC·NC·13·*	NC·DI·NC·2
			AO·NC·1	NP·NC·5·*	NP·DI·NC·2
		AQ·SP·NP·2	NP·NP·3·*	NC·AQ·NP·2	
M3	AO·1	NP·42·*	AQ·50·*	NC·135·*	
	AQ·NP·5	NP·NC·5·*	AQ·NC·22·*	NC·AQ·29·*	
	AO·NC·1	NP·NP·3·*		NC·NP·14·*	
	AQ·SP·NP·2	NP·AQ·1		NC·NC·13·*	
	AQ·DI·AQ·1	NC·NC·NC·5		NC·SP·NC·14·*	
	AQ·NP·NC·1	NC·AQ·NC·4		NC·SP·NP·7·*	
M4	AQ·50·*	NC·NP·14·*	AO·1	NC·135·*	NP·NC·NP·2
		NC·NC·13·*	AQ·NC·22·*	NP·42·*	NC·DI·NC·2
		NC·SP·NP·7·*	AQ·NP·5	NC·AQ·29·*	NP·DI·NC·2
		NC·NC·NC·5	AO·NC·1	NP·NC·5·*	NC·AQ·NP·2
			AQ·SP·NP·2	NP·NP·3·*	NC·NC·AQ·2
		AQ·NP·NC·1	NC·SP·NC·14·*	NC·NP·AQ·1	
M5	AQ·NC·22·*	NC·135·*	AQ·50·*	NP·42·*	NC·AQ·NC·4
		NC	AO·1	NC·AQ·29·*	NP·DI·AQ·2
		NC·14·*	AQ·NP·5	NC·NP·14·*	NP·NC·NP·2
			AO·NC·1	NC·NC·13·*	NC·DI·NC·2
			AQ·SP·NP·2	NP·NC·5·*	NP·DI·NC·2
		AQ·DI·AQ·1	NP·NP·3·*	NC·AQ·NP·2	

**Table 9.** Sample of relevant terms obtained in five domains worked.

07-Digital Communication	08-Crisis Communication	09-Television	10-Cinema	11-Audiovisual design
amenaza	información	clase	cine	mapping
difusión	filtración	cultural	año	edificio
fotografía	caso	social	industria	proyección
gigapíxeles	documento	cultura	público	técnica
imagen	periodismo	actividad	innovación	contenido
internet	fuentes	tipo	distribución	efecto
panorámica	secreto	consumo	exhibición	espacio
publicación	wikileaks	emoción	muerte	contexto
tecnología	vaticanleaks	fanático	tecnología	percepción
....	....	....	....	....

It should be noted that in the documents there are concepts of different lengths, one, two, three, four or more words; the tool removes some patterns that are rare and therefore are not taken into account, which diminishes the accuracy value, since these concepts, if they are not defined in the proposed patterns, will not be found.

### 4.5 Final Concepts Selection

To the list of candidate concepts, a parameter known as the “cut level” is applied, it is set at 70%, that is, we only keep the best 70% of the candidates, this level helps to discard some other concepts that may appear only once in the text, some of which may be relevant and others may not.

In addition to the original process, in this research was done a manually identification of each valid concept for the 11 documents of this study case, the results can be obtained by comparing those found by the expert and those found by the tool.

An example of the final concepts for each document can be seen in Table 10, in which it is possible to see that most are good patterns, however, patterns that are not very good are also included as in 04-Publicidad, where the pattern NP NP NP does not provide right concepts:

**Table 10.** Sample of final concepts of some domains.

01- Creativity	02- Social Media	03-Media Planning	04-Advertising
--AQ--	--NC--	--AQ--	--NC--
8·nuevo	7·manera	9·publicitario	9·pasado
2·fijo	6·clase	6·íntimo	7·nostalgia
2·itinerante	5·aprendizaje	4·público	--NP--
2·sorprendente	4·pared	3·líquido	3·polaroid
--NC--	--NP--	--NC--	2·ray_ban
7·espacio	3·mooc	10·medio	2·jameson
5·tienda	2·massive_open_on	4·lavabo	--AQ--
4·restaurante	line_courses	4·publicidad	3·actual
4·marca	2·internet	4·espacio	2·propio
--NP--	2·web	--NC AQ--	--NC AQ--
6·pop-up	2· --NC AQ--	3·medio publicitario	1·vínculo emocional
2·nueva_york	2·red social	2·mensaje publicitario	·público objetivo
2·londres	1·mundo profesional	2·medio íntimo	1·foto antiguo
2·barcelona	1·aulas virtuales	2·papel higiénico	--NC SP NC--
--NC AQ--	--NC SP NC--	--NC SP NC--	1·sensación de autenticidad
2·casa particular	1·aulas con pared	1·punto de vista	1·estilo de vida
1·experiencia	1·espacio con pared	1·publicidad de guerra	1·cantidad de dinero
irrepetible	1·clase sin pared	1·guerra en retretes	--NP NP NP--
1·prestigio	1·educación a distancia	1·cuarto de baño	1·ray_ban vinils mini
internacional	1·proceso de aprendizaje	1·muestra de perfume	1·vinils mini polaroid
1·entorno	aprendizaje	--NP--	1·mini polaroid tetris
sorprendente	.....	2·portugal	.....
.....		.....	

## 5 Results

Table 11 shows the results obtained for this study case, the 11 different domains considered were processed.

In this table the name of the documents are shown in the first column, the words per document are in the second column, the number of concepts found in the document

**Table 11.** Results obtained after processing the 11 different domains documents.

Document Name	Words per document	Originals	Candidates	Founds	Precision	Recall	F-Score
		#	#	#	%	%	%
01- Creativity	609	103	178	70	67.96	39.33	49.82
02- Social Media	764	121	154	71	58.68	<b>46.10</b>	51.64
03-Media Planning	834	142	247	83	58.45	33.60	42.67
04-Advertising	977	152	244	94	61.84	38.52	47.47
05-Journalism	755	68	173	48	<b>70.59</b>	27.75	39.83
06-Political Communication	727	82	201	52	63.41	25.87	36.75
07-Digital Communication	683	88	171	52	59.09	30.41	40.15
08-Crisis Communication	1094	185	274	122	65.95	44.53	<b>53.16</b>
09-Television	894	144	276	99	68.75	35.87	47.14
10-Cinema	1095	167	296	103	61.68	34.80	44.49
11-Audiovisual Design	735	120	190	79	65.83	41.58	50.97
<b>Average</b>					63.84	36.21	45.83

manually (originals) are in the third column, the number of candidate concepts (found by the proposed method) are in the fourth, the number of candidate concepts equal to the originals are in the fifth and in the following 3 are the parameters of Precision, Recall and F-Score, which help us to compare these results in a standard way.

The cells that are above the general average of the 11 documents were marked in gray, for Precision and Recall 5 of 11 are above average and for the F-Score column, 6 of 11 elements are above the average, in bold the highest values of each column are found, it is highlighted that 8 of the 11 are above 60% of Precision.

## 6 Conclusions

The results of the research show that it is possible to extract good key concepts from a particular domain from a single document (in this case 11 different areas were studied) without the necessity of a large corpus, it is concluded that it is possible to identify up to 70% of the validated concepts in the documents with a maximum precision of 46%, improving the maximum 25% of [1]. The key elements for this investigation were: Freeing, the automatic pattern learning, the modification to filter out undesirable/unproductive patterns and the modification to detect single word patterns.

Consequently, these helped us to obtain the important concepts for each document. The adaptation that was made to be able to obtain the formed concepts by a single word, helped to conform 70% of our valid concepts, since without this adaptation, 82% of the original concepts of the document would be lost.

We also found potential to improve the automatic pattern learning process adding neural networks or genetic algorithms, which will be presented in a new publication.

## References

1. Ochoa, J.L., Almela, A., Hernández-Alcaraz, M.L., Valencia-García, R.: Learning morphosyntactic patterns for multiword term extraction. *Sci. Res. Essays* **6**(26), 5563–5578 (2011)
2. Harris, Z.S.: Distributional structure. *WORD* **10**(2–3), 146–162 (1954). <https://doi.org/10.1080/00437956.1954.11659520>
3. Elnagar, A., Einea, O., Lulu, L.: Comparative study of sentiment classification for automated translated Latin reviews into Arabic. In: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), Hammamet, Tunisia, pp. 443–448 (2018). <https://doi.org/10.1109/aiccsa.2017.82>
4. Wolk, K., Glinkowski, W., Żukowska, A.: Enhancing the assessment of (Polish) translation in PROMIS using statistical, semantic, and neural network metrics. In: Rocha, Á., Adeli, H., Reis, L.P., Costanzo, S. (eds.) *WorldCIST'18 2018*. AISC, vol. 746, pp. 351–366. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-77712-2\\_34](https://doi.org/10.1007/978-3-319-77712-2_34)
5. Vani, K., Gupta, D.: Text plagiarism classification using syntax based linguistic features. *Expert Syst. Appl.* **88**, 448–464 (2017). <https://doi.org/10.1016/j.eswa.2017.07.006>
6. Vani, K., Gupta, D.: Detection of idea plagiarism using syntax–semantic concept extractions with genetic algorithm. *Exp. Syst. Appl.* **73**, 11–26 (2017). <https://doi.org/10.1016/j.eswa.2016.12.022>
7. Vani, K., Gupta, D.: Unmasking text plagiarism using syntactic-semantic based natural language processing techniques: comparisons, analysis and challenges. *Inf. Process. Manag.* **54**(3), 408–432 (2018). <https://doi.org/10.1016/j.ipm.2018.01.008>
8. Moradi, M., Ghadiri, N.: Different approaches for identifying important concepts in probabilistic biomedical text summarization. *Artif. Intell. Med.* **84**, 101–116 (2018)
9. Yousefi-Azar, M., Hamey, L.: Text summarization using unsupervised deep learning. *Exp. Syst. Appl.* **68**, 93–105 (2017). <https://doi.org/10.1016/j.eswa.2016.10.017>
10. Gambhir, M., Gupta, V.: Recent automatic text summarization techniques: a survey. *Artif. Intell. Rev.* **47**(1), 1–66 (2017). <https://doi.org/10.1007/s10462-016-9475-9>
11. Spasić, I.: Acronyms as an integral part of multi-word term recognition—a token of appreciation. *IEEE Access* **6**, 8351–8363 (2018). <https://doi.org/10.1109/ACCESS.2018.2807122>
12. Ali, A.A., Saad, S.: Unsupervised concept hierarchy induction based on Islamic glossary. *ARPN J. Eng. Appl. Sci.* **11**(13), 8505–8510 (2016)
13. Ochoa, J.L., Hernandez-Alcaraz, M.L., Almela, A., Valencia-Garcia, R.: Learning semantic relations from Spanish natural language documents in the financial domain. In: *IEEE 3rd International Conference on Computer Modeling and Simulation (ICCMS 2011)*, Mumbai, India (2011)
14. Kuang, Z., Yu, J., Li, Z., Zhang, B., Fan, J.: Integrating multi-level deep learning and concept ontology for large-scale visual recognition. *Pattern Recognit.* **78**, 198–214 (2018). <https://doi.org/10.1016/j.patcog.2018.01.027>
15. Ochieng, P., Kyanda, S.: Ontologies' mappings validation and annotation enrichment through tagging. *Artif. Intell. Rev.* 1–28 (2018). <https://doi.org/10.1007/s10462-018-9632-4>
16. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. *Sci. Am.* **284**(5), 34–43 (2001)
17. Maedche, A., Staab, S.: Ontology learning for the Semantic Web. *IEEE Intell. Syst.* **16**(2), 72–79 (2001). <https://doi.org/10.1109/5254.920602>
18. Unadkat, R.: Survey paper on semantic web. *Int. J. Adv. Pervasive Ubiquit. Comput. (IJAPUC)* **7**(4), 13–17 (2015). <https://doi.org/10.4018/IJAPUC.2015100102>

19. Chen, P.H., Zafar, H., Galperin-Aizenberg, M., Cook, T.: Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports. *J. Digit. Imaging* **31**(2), 178–184 (2018). <https://doi.org/10.1007/s10278-017-0027-x>
20. Nguyen, A., Lawley, M., Hansen, D., Colquist, S.: Structured pathology reporting for cancer from free text: Lung cancer case study. *Electr. J. Health Inf.* **7**(1) (2012). Art. No. e8
21. Wermter, J., Hahn, U.: Finding new terminology in very large corpora. In: Proceedings of the 3rd International Conference on Knowledge Capture (K-CAP 2005), pp. 137–144. ACM, New York (2005). <http://dx.doi.org/10.1145/1088622.1088648>
22. Zouaq, A., Nkambou, R.: Enhancing learning objects with an ontology-based memory. *J. IEEE Trans. Knowl. Data Eng.* **21**(6), 881–893 (2009)
23. Ochoa, J.L., Almela, A., Ruiz-Martínez, J.M., Valencia-García, R.: Efficient multiword term extraction in Spanish. Application to the financial domain. In: IEEE International Conference on Intelligence and Information Technology (ICIIT 2010), Lahore, Pakistan, vol. 1, pp. 426–430 (2010)
24. Frantzi, K.T., Ananiadou, S., Tsujii, J.: The *C-value/NC-value* method of automatic recognition for multi-word terms. In: Nikolaou, C., Stephanidis, C. (eds.) ECDL 1998. LNCS, vol. 1513, pp. 585–604. Springer, Heidelberg (1998). [https://doi.org/10.1007/3-540-49653-X\\_35](https://doi.org/10.1007/3-540-49653-X_35)
25. Frantzi, K.T., Ananiadou, S.: The *C-value/NC value* domain independent method for multi-word term extraction. *J. Nat. Lang. Process.* **3**(6), 145–180 (1999)
26. Frantzi, K.T., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms: the *C-value/NC-value* method. *Int. J. Digit. Libr.* **3**(2), 115–130 (2000)
27. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **24**(5), 513–523 (1988)
28. Hisamitsu, T., Tsujii, J.: Measuring term representativeness. In: Paziienza, M.T. (ed.) Information Extraction in the Web Era. LNCS (LNAI), vol. 2700, pp. 45–76. Springer, Heidelberg (2003). [https://doi.org/10.1007/978-3-540-45092-4\\_3](https://doi.org/10.1007/978-3-540-45092-4_3)
29. Paziienza, M.T., Pennacchiotti, M., Zanzotto, F.M.: Terminology extraction: an analysis of linguistic and statistical approaches. In: Sirmakessis, S. (ed.) Knowledge Mining. Studies in Fuzziness and Soft Computing, vol. 185, pp. 255–279. Springer, Heidelberg (2005). [https://doi.org/10.1007/3-540-32394-5\\_20](https://doi.org/10.1007/3-540-32394-5_20)
30. Rousseau, F., Vazirgiannis, M.: Main core retention on graph-of-words for single-document keyword extraction. In: Hanbury, A., Kazai, G., Rauber, A., Fuhr, N. (eds.) ECIR 2015. LNCS, vol. 9022, pp. 382–393. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-16354-3\\_42](https://doi.org/10.1007/978-3-319-16354-3_42)
31. Yao, X., Gan, J., Xu, J.: Concept extraction based on hybrid approach combined with semantic analysis. In: 2017 International Conference on Applied Mechanics and Mechanical Automation (AMMA 2017)
32. Smatana, M., Butka, P.: Extraction of keyphrases from single document based on hierarchical concepts. In: IEEE 14th International Symposium on Applied Machine Intelligence and Informatics (SAMi), Herlany, pp. 93–98 (2016). <https://doi.org/10.1109/sami.2016.7422988>
33. Al-Abdallah, R.Z., Al-Taani, A.T.: Arabic single-document text summarization using particle swarm optimization algorithm. *Proc. Comput. Sci.* **117**, 30–37 (2017). <https://doi.org/10.1016/j.procs.2017.10.091>
34. Padró, L., Stanilovsky, E.: FreeLing 3.0: towards wider multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul, Turkey. (2012). UPCommons, <http://hdl.handle.net/2117/15986>. Accessed 08 June 2018

35. Ochoa, J.L., Hernández-Alcaraz, M.L., Valencia-García, R., Martínez-Béjar, R.: A semantic role-based methodology for knowledge acquisition from Spanish documents. *Int. J. Phys. Sci.* **6**(7), 1755–1765 (2011)
36. Subramaniam, T., Jalab, H.A., Taga, A.Y.: Overview of textual antispam filtering techniques. *Int. J. Phys. Sci.* **5**(12), 1869–1882 (2010)
37. Kotelnikov, E., Razova, E., Fishcheva, I.: A close look at russian morphological parsers: which one is the best? In: Filchenkov, A., Pivovarova, L., Žižka, J. (eds.) *AINL 2017. CCIS*, vol. 789, pp. 131–142. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-71746-3\\_12](https://doi.org/10.1007/978-3-319-71746-3_12)
38. Ferrés, D., AbuRa'ed, A., Saggion, H.: Spanish morphological generation with wide-coverage lexicons and decision trees. *Procesamiento del Lenguaje Natural*, SI, **58**, 109–116 (2017)
39. Vázquez-García, M.: (COMeIN) Disección de contenidos a partir de conceptos. <http://comein.uoc.edu/divulgacio/comein/es/numero15/articles/Article-Merce-Vazquez.html>. Accessed June 08 2018