# Finding Answers to Definition Questions across the Spanish Web

Alejandro Figueroa
German Centre for Artificial Intelligence - DFKI
Stuhlsatzenhausweg 3, D - 66123, Saarbrücken, Germany
figueroa@dfki.de

## ABSTRACT

In the last two years, we have been developing a multilingual web question answering system, which is aimed at discovering answers to natural language questions in three different languages: English, German and Spanish. One of its major components is the module that extracts answers to definition questions from the web.

This paper compares and provides insights into different techniques that we have used during these two years for tackling definition questions in Spanish. Additionally, the present work focuses its attention on new challenging issues regarding the language phenomena that adversely affect the answering process. In particular, in comparison with the analogous task in English.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.7.m [**Document and Text Processing**]: Miscellaneous

## General Terms

Algorithm, Languages, Measurement

## Keywords

Web search, search for definitions, text mining, web mining, web question answering, Spanish, definitions in Spanish.

## 1. INTRODUCTION

Web users are constantly seeking information about concepts, persons, locations or things in their daily lives. This "*tell me about*" task is an inherent part of being human, and in past decades, people looked for the required information in dictionaries and encyclopedias on their bookshelves or well-stocked libraries nearby. Nowadays, commercial search engines make it possible to find information about wide-ranging topics across a massive and constantly changing collection of documents, called the web.

Essentially, definition questions, including "*Who is Rafael Nadal?*" and "*What is www09?*", match this "*find out about*" requirement. In recent years, this type of question has become especially interesting due to their high frequency in real user logs. As often as not, an answer to a definition question consists essentially of a set of nuggets that convey correct, precise and succinct information about the topic of the question (a.k.a. *definiendum* or target). In the jargon of definition questions, a nugget is a piece of relevant or factual information about the *definiendum*. In general, short sentences are seen as nuggets because they ensure the necessary context to understand their meaning.

However, determining this set of nuggets is an extremely difficult task, and it usually involves the next three steps:

* A search strategy that substantially boosts the recall of promising documents.
* An answer extraction method that recognises correct and reliable answers.
* A summarisation module that removes redundant information from the final output.

Broadly speaking, Question Answering Systems (QAS) that automatically answers definition questions have been widely studied in the context of the question answering track of the Text REtrieval Conference (TREC). While TREC is aimed specifically at English, the Cross-Language Evaluation Forum (CLEF) motivates the research of QAS operating on European languages, such as Spanish. In CLEF, QAS are aimed at the EFE corpus, which is comprised of about 450,000 documents.

Contrary to CLEF systems, one of the components of our system (`EsDefWebQA`) extracts definitions from the Spanish web. This work compares three strategies, different in nature, utilised by `EsDefWebQA`. Additionally, this paper analyses the differences in the underlying linguistic phenomena that make answering this type of question in Spanish more difficult than in English in depth.

The organization of this paper is as follows: section 2 discusses the related work, section 3 describes our search strategy, section 4 presents our answer extraction methods, section 5 shows the results obtained by applying our approaches and finally section 6 highlights the main conclusions and further work.

## 2. RELATED WORK

Some systems in TREC recognise descriptive sentences by making use of lexico-syntactic patterns that often convey definitions. These patterns are normally applied at the surface [17] or constituent level [12], and they basically include clues such as "*is/are a/an/the*"[1]. The main drawback of

[1]The reader can also refer to [10] for a more extensive list of clues.


*WWW2009*, April 20-24, 2009, Madrid, Spain.
.

1: <definiendum> [es|son|fueron|fue|ha sido|han sido] [la|lo|el|un|una|uno|unos|unas|las|los] <description>
e.g.,"***Marcelo Bielsa*** *es el nuevo seleccionador de fútbol de Chile."*
2: <definiendum> [,|;] [un|una|uno|la|lo|el|los|las] <description> [,|;|.]
e.g.,"***Google****, el buscador más usado de internet, fijará su sede en Valencia... "*
3: <definiendum> [ha llegado a ser|llego a ser|se transformo|se ha transformado] <description>
e.g.,"*El* ***Euro*** *se ha transformado en una de las monedas más fuertes del planeta.*"
4: <definiendum> [,|] [el cual|la cual|los cuales|quien|que] <description>
e.g.,"***Neil Armstrong***, quien fue el primer hombre en la Luna."
5: <definiendum> [nacio|fue fundado|fue fundada] <description>
e.g.,"***Rafael Nadal*** nacio el 3 de Junio de 1986 en Manacor, Mallorca."

**Table 1: Surface Patterns for Identifying Definitions in Spanish.**

these patterns is that they fail to provide an unerring accuracy. That is, not all sentences matching these patterns necessarily express descriptive information. In the same way, the best system (INAOE) in the Spanish track of CLEF 2006 took advantage of lexico-syntactic patterns for identifying answers to definition questions in Spanish [3]. The INAOE system learnt these patterns from sentences extracted from the web [1, 8]. By large, their patterns are grounded on: (a) the position of the definition with respect to the *definiendum*, and (b) stop-words and punctuation. To neatly illustrate these patterns, consider the following two examples:

```
, el <description>, <definiendum>, dijo
y el <description>, <definiendum>.
```

In our previous work [10], this set of patterns was extended by taking into account translations of the English lexico-syntactic constructs into Spanish. These translations are sketched in table 1, and they assist in overcoming some special difficulties presented when taking advantage of the patterns by [1, 8]. As a result, these patterns helped our system to answer 22 out of 35 questions, for which the CLEF 2006 gold standard supplied an answer nugget. It is fair to highlight here that these answers were distinguished on the web, contrary to CLEF systems.

In order to identify correct answers across the AQUAINT corpus, QAS in TREC take nuggets from several external specific resources of descriptive information (e.g. online encyclopedia and dictionaries), and project them into a set of paragraphs/sentences retrieved from the corpus afterwards. Generally speaking, this projection strategy relies heavily on: (a) finding entries, corresponding to the *definiendum*, in these external resources, and (b) finding a significant overlap between terms in definitions within the target collection and those taken from these specific resources.

For instance, reference [6] took advantage of external resources like WordNet glossaries, online specific resources (e.g., Wikipedia) and web snippets for learning frequencies of words that correlate to the *definiendum*. One of their findings was that definitional web-sites greatly enhance the performance, leading to few unanswered questions: Wikipedia covered 34 out of the 50 TREC–2003 definition queries and biography.com 23 out of 30 questions regarding people, all together providing answers to 42 queries. Afterwards, reference [6] made use of these correlated words for forming a centroid vector. Sentences were thereafter ranked according to the cosine distance to this vector.

By the same token, the best system in the Spanish CLEF 2007 track followed this dominant trend. This system looked the *definiendum* up in Wikipedia articles in Spanish, and projected the words within the first descriptive line into a set of selected passages taken from the EFE corpus [2, 7].

The beneficial aspect of strategies grounded on the overlapping of strongly correlated words is that they are based on the Distributional Hypothesis [11]. This means they rank answer candidates according to the degree in which their respective words characterise the *definiendum*. There are, on the other hand, two essential aspects that make these strategies less attractive: (a) they normally fail to recognise right answers (sentences) with words that observe a low correlation with the *definiendum* across the specific resources, causing a less diverse output and a detriment to coverage, and more important, (b) taking into account only semantic relationships is insufficient for ranking answer candidates. Contrary to external resources, the co-occurrence of the *definiendum* with learnt words across sentence/answer candidates does not necessarily guarantee that they are syntactically dependent (see for example [4]). Consider the following illustrative example regarding "*El primer ministro británico Gordon Brown*":

```
El primer ministro, François Fillon, ha dirigido, el
miércoles 27 de junio, en nombre del gobierno francés,
una carta de enhorabuena a Gordon Brown, ....
```

Another reason that makes this type of approach less attractive is that the coverage provided by specific external resources, including Wikipedia, widely varies from English to another language. For instance, at the time of writing, Wikipedia supplied about 2,500,000 English articles, whereas only about 450,000 articles in Spanish. It is thus crystal clear that this sort of projection strategy will leave several unanswered or incompletely answered questions.

With regards to the completeness of the answer, the assessment of definition QAS radically differs from TREC to CLEF. In CLEF, systems are encouraged to provide a short description of the *definiendum*. This definition is then assessed as right or wrong regardless of its length and how informative it is. In TREC, conversely, systems are rewarded for detecting as much diverse descriptive information as possible, while at the same time, they are penalised for long and redundant answers [18].

## This work

This work is an amalgamation of a detailed analysis of the intrinsic factors, including linguistic phenomena, that mit-

$q_1$:"<definiendum>"
$q_2$:"<definiendum>, fue un" ∨ "<definiendum> son lo" ∨ "<definiendum>, la"
$q_3$:"<definiendum> fue la" ∨ "<definiendum> es el" ∨ "<definiendum> son el"
$q_4$:"<definiendum> que" ∨ "<definiendum> son las" ∨ "<definiendum>, lo"
$q_5$:"<definiendum> es un" ∨ "<definiendum> ha llegado a ser" ∨ "<definiendum> son la" ∨ "<definiendum> fueron las"
$q_6$:"<definiendum> fue el" ∨ "<definiendum> son unas" ∨ "<definiendum>, uno" ∨ "<definiendum> ha sido la"
$q_7$:"<definiendum> quien" ∨ "<definiendum> los cuales" ∨ "<definiendum>, un" ∨ "<definiendum> son una"
$q_8$:"<definiendum> se ha transformado" ∨ "<definiendum> es lo" ∨ "<definiendum> fue fundado"
$q_9$:"<definiendum>, el" ∨ "<definiendum> son unos" ∨ "<definiendum> fue una" ∨ "<definiendum> fue fundada"
$q_{10}$:"<definiendum> es la" ∨ "<definiendum> llego a ser" ∨ "<definiendum> ha sido el" ∨ "<definiendum> son un"
$q_{11}$:"<definiendum> es una" ∨ "<definiendum> fue lo" ∨ "<definiendum> ha sido un"
$q_{12}$:"<definiendum> se transformo" ∨ "<definiendum> fue uno" ∨ "<definiendum> , las"
$q_{13}$:"<definiendum> la cual" ∨ "<definiendum>, una" ∨ "<definiendum> ha sido una"
$q_{14}$:"<definiendum> es uno" ∨ "<definiendum> nacio" ∨ "<definiendum> el cual" ∨ "<definiendum>, los"

**Table 2: Queries for Searching Definitions in Spanish.**

igate the performance of web QAS in Spanish, and a comparison between three strategies, used by `EsDefWebQA`, in consonant with the TREC evaluation methodology.

## 3. SEARCHING FOR DEFINITIONS

As it was briefly described in [10], our search strategy has not changed that much. In order to make all posterior sections clearer, this section fleshes this strategy out, and it greatly extends our remarks concerning the underlying factors involved in the search process.

Table 1 lists a set of surface patterns that often indicate definitions in Spanish. `EsDefWebQA` takes advantage of these constructs not only for distinguishing definitions within documents, but also for biasing the search engine in favour of web snippets, and thus documents, that are very likely to convey descriptions of the *definiendum*. This bias is achieved by automatically generating and submitting fourteen search queries (sketched in table 2), where their clauses are based on the surface patterns listed on table 1. Plainly speaking, the more successful this query rewriting is, the larger the recall of web snippets, and hence documents, containing definitions is.

The reason why a noticeable increase in the recall of descriptive information enhances the chance of correctly answering a definition question is two-fold: (a) it increases the probability of matching the context of a model previously learnt from annotated examples, including those taken from online encyclopedias and dictionaries, and consequently (b) it makes the selection of the most relevant and reliable, as well as descriptive answers, easier.

Certainly, the success of this strategy lies in the size of the target corpus, in this case, the Spanish web. A larger corpus tends to provide a wider coverage, and therefore, likely to assist QAS in leaving less unanswered questions. But more important, a considerably larger corpus yields a massive redundancy. It is worth duly pointing out that, by redundancy it is not meant duplicate information, but rather different paraphrases of the same underlying ideas. QAS can undoubtedly benefit from paraphrases, because they markedly increase the probability of matching query terms and purpose-built patterns. Consequently, they considerably boost the chance of finding more and fuller answers.

Unfortunately, there is a big difference between the number of web documents in English and Spanish. As a very rough rule of thumb, we estimate this difference approximately by submitting some lexico-syntactic clues to the web in order to get their web frequency counts. Table 3 emphasises this difference.

| English | | Spanish | |
|---|---|---|---|
| is the | 6,840,000,000 | es un | 323,000,000 |
| is a | 8,720,000,000 | es una | 172,000,000 |
| is an | 2,440,000,000 | es la | 162,000,000 |
| | | es el | 150,000,000 |
| | | es uno | 36,700,000 |
| | + | es las | +1,710,000 |
| | 18,000,000,000 | | 845,410,000 |

**Table 3: Definition Clues Frequencies Comparison.**

Our rough estimates indicate that the size of this corpus falls into a drastic decline from about 18 billion in English to 1 billion in Spanish. This comparatively small number of matches enforces `EsDefWebQA` to divert time and effort away from extracting answers to perform an exhaustive search for promising documents. In general, when answering definitions questions in English, few queries aimed at a small set of lexico-syntactic constructs suffice to obtain a high recall of web snippets, and hence, documents that contain descriptive information about the target concept. Since the amount of Spanish web documents is much smaller, the probability of matching these lexico-syntactic constructs dramatically decreases. Our system is, for this reason, compelled to submit a larger number of queries to the search engine, in order to sharply increase the probability of obtaining diverse and sufficient information to satisfactorily answer the question.

There are also linguistic aspects that make the search process more demanding. Most nouns in modern English lack grammatical gender. This gender is triggered by two indefinite articles: "a" and "an", and one definitive article: "the", which is also used for indicating plural forms. Therefore, a query as follows would be enough to retrieve most of the describing nouns:

$q_*$:"<definiendum> is a" ∨ "<definiendum> is an" ∨ "<definiendum> is the"

Conversely, Spanish uses three grammatical genders: feminine, masculine, and neuter, which are signalled by six definite and indefinite articles. Furthermore, Spanish utilises four additional morphological forms of these articles for agreeing the number of the noun phrase they modify, that is indicating plural nouns. All together, this increases the number of articles from three to ten. The reader can check increase by inspecting the first pattern in table 1.

This growth in linguistic complexity brings about an extra effort that goes into the search process. More specifically, a richer noun morphology leads to more lexico-syntactic clues, which means more search clauses, and by the same token, a longer retrieval time.

On the whole, our system submits the fourteen queries shown in table 2. Each of them is aimed at fetching thirty web snippets.

## 3.1 Filtering Out Unpromising Answer Candidates

As mentioned in section 2, definition surface patterns do not supply a pinpoint accuracy. Indeed, these constructs can be used for conveying several types of information, including opinions. But more important, the fact that a sentence matches a pattern does not necessarily mean that it conveys descriptive information about *definiendum*. To neatly illustrate this point, consider the following example than was fetched when searching for "*Hugo Chávez es la*":

```
Una de las reformas más importantes de Hugo Chávez es la
constitución que un Fondo Monetario Latinoamericano, que
llaman Banco del Sur y que precisamente en ...
```

This illustrative example communicates information about a reform advocated by *Hugo Chávez*, but not about himself. Regularly, the matched *definiendum* ($\delta_m$) does not exactly match the input of the user ($\delta_u$). `EsDefWebQA` takes advantage of the *Jaccard Measure* for distinguishing more reliable descriptive sentences, and in like manner, for improving the accuracy of the pattern matching strategy. The *Jaccard Measure* $J$ of the two *definiendum* $\delta_u$, $\delta_m$ is the ratio between the number of different unigrams that they share, and the total number of different unigrams:

$$J(\delta_u, \delta_m) = \frac{\mid \delta_u \cap \delta_m \mid}{\mid \delta_u \cup \delta_m \mid}$$

In our working example, the *Jaccard Measure* between "*Hugo Chávez*" and "*Una de las reformas más importantes de Hugo Chávez*" is $\frac{2}{8} = 0.25$. `EsDefWebQA` filters reliable descriptive utterances by means of a pattern specific threshold. This way it avoids additional purpose-built hand-crafted rules and ad-hoc linguistic processing [17]. The experimental value of these thresholds are 0.33 and 0.4, for the first and the last patterns in table 2, respectively. All remaining thresholds were experimentally set to 0.25.

The special advantage of this word overlapping methodology is that it can be applied to different languages indistinctly, which is vitally important in the design of multilingual QAS. However, applying this strategy to a new language inevitably involves computing new experimental thresholds. Still yet, there are two additional problems that arise when applying this strategy to Spanish: (a) the discarded sentences can possibly contain descriptive information that is not present in the group of sentences seen as reliable, and (b) sentences in Spanish do not necessarily need to contain an explicit subject. In the case of English, the former is considerably alleviated by the amount of redundancy provided by the web (see section 3). In order to explain the latter, let us consider the following first four sentences taken from the Wikipedia article regarding "*Genovevo Rivas Guillén*":

```
(1) Gral.  Genovevo Rivas Guillén (1886-1947) fue un
militar y Gobernador provisional de San Luis Potosí
mexicano.  (2) Nacio en Rayon, San Luis Potosí, en 1886.
(3) Lucho como maderista desde 1910, bajo las ordenes
del Gral.  Alberto Carrera Torres.  (4) Durante la
Expedición Punitiva se distinguió en la Batalla de
Carrizal, que fue un enfrentamiento contra tropas
norteamericanas que perseguían Francisco Villa en el año
1924, concediéndosele la condecoración del Valor Heroico.
```

In this paragraph, sentences 2-4 omit all explicit references to "*Genovevo Rivas*", but they nevertheless yield factual information about him. Sentence 4 especially serves to highlight the case of the passive "*se*" construction, which is chiefly used in third person, increasing its probability of being utilised for defining concepts.

The absence of references force QAS to process the entire paragraph, in order to determine to whom each sentence refers. While it is arguable that, in the case of biographical sources, the title and the position of the sentences are good features to solve this problem, it is also true that many other classes of documents do not observe these patterns, and they still yield descriptive information. Taking into account all sorts of documents is particularly important for languages where a small redundancy is provided. For instance, consider the following blog entry:

```
(1) La persona a quien más admiro es Ricky Martin.
(2) Es cantante.  (3) Nació en Puerto Rico en 1971.
(4) A la edad de seis años apareció en anuncios en la
televisión.  (5) Fue selecionado para el grupo "Menudo" a
los doce años.  (6) Con su primer álbum obtuvo ocho
discos de oro en México, Chile, Argentina, Puerto Rico y
Estados Unidos.
```

In this blog entry, the *definiendum* is the direct object of first sentence, and all the posterior sentences talk about this object without being explicitly referenced. Conversely, descriptive sentences in English usually convey an explicit subject, making it easier to disambiguate about who/what they are talking. In the particular case of definitions, the subject can refer to the *definiendum* by means of pronouns (e.g. he, it, her ), orthographical variations, aliases or synonyms.

All things considered, recognising implicit subject pronouns is particularly important to maximise the chances of identifying descriptive information from different types of documents as much as possible. This linguistic phenomenon reaffirms the need for a higher level of redundancy, and from our standpoint, it stresses the need for deeper linguistic processing at the paragraph level.

# 4. EXTRACTING AND SUMMARISING ANSWERS

`EsDefWebQA` is aimed at selecting a subset of the reliable

sentences determined in the previous steps. This selection is aimed basically at maximising the diversity and reducing the redundancy of the final output. The following sections deal at length with three different strategies of extracting answers to definition questions, utilised by `EsDefWebQA`.

## 4.1 Strategy One

Since `EsDefWebQA` is part of our multilingual question answering system, our first approach was aimed at being unsupervised and largely language independent. The basic assumption of our strategy is that terms high in frequency across candidate sentences are very likely to signal reliable answers [13, 14]. Here, reliability goes hand-in-hand with representativeness, that is candidate sentences strongly overlapping with other candidate sentences are more delineative. But more important, this assertion implicitly means that some definitions can be inferred directly from their contextual evidence, that is without the assistance of an oracle of descriptive information, like online encyclopedias and dictionaries. This is very relevant when tackling definition questions in languages different than English, because these specific external resources yield narrow coverage. Hence, in these languages, accounting for sentences taken from different types of web documents offers a workable solution. Accordingly, the coverage of a sentence $S_s \in S$ is defined as:

$$coverage(S_s, \phi) = \sum_{\forall w_i \in W_{S_s} - \phi} P(w_i)$$

Where $W_{S_s}$ is the set of all words in the reliable sentence $S_s$, and $P(w_i)$ is defined as the probability of finding the word $w_i$ in the set of all reliable sentences $S$ identified in section 3.1. It is worth remarking here that $P(w_i)$ was arbitrarily set to zero for stop-words and, for the sake of clarity, $\phi$ is defined later. In this strategy, a list of stop-words is the only external knowledge used.

On the other hand, the definition content of a sentence is defined by the degree in which this sentence describes different aspects of the *definiendum*. That is, not all sentences convey the same amount of descriptive information. `EsDefWebQA` assesses the descriptive content of a sentence by accounting for two factors: (a) sentences containing a larger number of entities are more likely to carry more descriptive information, because they presumably establish a relation between the *definiendum* and a larger number entities, and (b) the identification of a set of words that are likely to describe the different facets of the *definiendum*. `EsDefWebQA` determines this set of words by inspecting the forty closest neighbours to the *definiendum* in the semantic space provided by *Latent Semantic Analysis* (LSA). In this semantic space, the neighbourhood of a particular word provides its context [15]. Consequently, the definition content of a sentence in $S$ is defined as follows:

$$content(S_s, \phi) = \sum_{\forall w_i \in W_{S_s} \cap (\bar{W} \cup \phi)} R(w_i) + \sum_{\forall e_e \in S_s - \phi} P(e_e)$$

Where $P(e_e)$ is the probability of finding the entity $e_e$ in $S$. Since this approach is aimed at achieving a high degree of language independency, these entities are discriminated on the ground of sequences of words that start with a capital letter or a number. $R(w_i)$ is the degree of semantic relation between $w_i$ and the *definiendum* supplied by LSA. This semantic relation is only valid for the set $\bar{W}$ of the forty closest neighbours to the *definiendum*.

**Algorithm 1** Strategy One

```
1: φ = ∅;
2: while true do
3:    nextSen = null
4:    for all S_s ∈ S do
5:       rank = rank(S_s,φ);
6:       if nextSen == null or rank > rank(nextSen) then
7:          nextSen = S_s
8:       end if
9:    end for
10:   if nextSen == null or rank(nextSen) ≤ λ then
11:      break;
12:   end if
13:   print nextSen
14:   add(nextSen,φ)
15: end while
```

Eventually, the rank of a sentence $S_s$ is given by the sum of its coverage and content. `EsDefWebQA` discovers answers to definition questions by iteratively ranking and selecting reliable sentences $S$. Algorithm 1 shows this iterative strategy, where its input is the set $S$.

`EsDefWebQA` is not aimed only at incrementally selecting sentences that convey definitions, but also at reducing the size of this set by lessening the amount of redundancy. For this purpose, `EsDefWebQA` initialises a set $\phi$, in which it stores words and entities belonging to previously selected sentences (line 1). Next, sentences are ranked according to their coverage and content (line 5). However, when ranking, `EsDefWebQA` takes into consideration $\phi$. $\phi$ assists `EsDefWebQA` in ranking sentences according to their novelty with respect to previously selected sentences, while at the same time, $\phi$ causes sentences carrying redundant information to systematically decrease their ranking value. Here, the idea of redundancy is in consonant with the overlap between the word/entity and $\phi$. Then, `EsDefWebQA` chooses the highest ranked sentence after each iteration (line 6-8), and its corresponding words and entities are added to $\phi$ (line 14). If the highest ranked sentence meets the conditions in line 10, the extraction process finishes (line 11). These conditions include: (a) there is no more sentences to select, or (b) there are no more candidate sentences that show strong evidence of having novel and reliable descriptive information, that is all non-selected sentences have a ranking value lower than an experimental threshold $\lambda$=0.1.

In brief, in this answer extraction approach, candidate sentences become less attractive as long as their overlap with **all** previously selected sentences become larger. Contrary to other approaches for English that control the overlap at the word level [4, 12]. For example, [12] discarded a random sentence from each pair that shared more than 60% of their words. Conversely, `EsDefWebQA` can still select a sentence that substantially overlap if the novel content is very likely to convey a description. Additionally, in approaches like [12], a selected sentence can still be overlapped with a group of selected sentences. Unlike [12], [19] took advantage of the edit distance, and [4] made use of the cosine similarity between the new selected sentence and the previously selected sentences, and whenever this similarity was greater than a threshold, the new sentence was discarded. By large,

| $\vec{w} =< w_1, w_2 >$ | $I_2(\vec{w})$ | $\vec{w} =< w_1, w_2 >$ | $I_2(\vec{w})$ | $\vec{w} =< w_1, w_2, w_3 >$ | $I_3(\vec{w})$ |
|---|---|---|---|---|---|
| <cientista, $w_*$ > | 10.54 | < $w_*$, analista > | 8.70 | < $w_*$, democristiano, de> | 28.16 |
| < $w_*$, democristiano> | 9.75 | < $w_*$, derechista> | 8.66 | < $w_*$, catalanista, en> | 27.72 |
| < $w_*$, socialdemocrata> | 9.52 | < $w_*$, federalista> | 8.64 | < $w_*$, centrista, de> | 27.16 |
| < $w_*$, democratacristiano> | 9.50 | < $w_*$, regionalista> | 8.64 | < $w_*$, aboga, por> | 26.53 |
| < $w_*$, afiliado> | 9.43 | <partido, $w_*$ > | 8.57 | < $w_*$, centro-izquierda, de> | 26.26 |
| < $w_*$, trotskista> | 9.34 | < $w_*$, salvadoreño> | 8.50 | < $w_*$, afiliado, al> | 26.05 |
| < $w_*$, catalanista> | 9.30 | < $w_*$, marxista-leninista> | 8.49 | < $w_*$, centrista, fundado> | 25.99 |
| < $w_*$, centro-derecha> | 9.17 | < $w_*$. intendente> | 8.47 | < $w_*$, socialdemocrata, de> | 25,92 |
| < $w_*$, centro-izquierda> | 9.09 | < $w_*$. sindicalista> | 8.43 | < $w_*$, centro-derecha, de> | 25.75 |
| < $w_*$, galleguista> | 9.00 | <activismo. $w_*$ > | 8.43 | < $w_*$, sindicalista, espaol> | 23.43 |
| <abogado. $w_*$ > | 8.89 | < $w_*$, democrata> | 8.30 | < $w_*$, independentista, vasca> | 22.55 |

**Table 4: Some Strong Word Association Norms with $w_*$="politico".**

[12] and [4]'s strategies suffer from the same drawbacks.

## 4.2 Strategy Two

The previous answer extraction method suffers from the following drawback: many candidate sentences are descriptive, but they do not significantly overlap with other sentences, obtaining a low coverage and entity content. This second method is aimed at tackling this problem head-on.

The underlying idea behind our solution is learning lexico-syntactic regularities that are normally presented within descriptions. More precisely, these regularities are extracted from sentences taken from all abstracts in Wikipedia articles that match patterns in table 1. Eventually, when answering a definition question prompted by the user, sentences in $S$ matching these regularities increase their chances of being incorporated into the final output, even though they do not significantly overlap with other sentences in $S$.

This learning strategy is in contrast to the current trend, because `EsDefWebQA` does not project into the target set of sentences $S$, words corresponding to entries of the *definiendum* in online encyclopedias or dictionaries. But rather, `EsDefWebQA` learns these regularities from all articles, and hence any definition can assist in deciding whether or not a candidate sentence in $S$ expresses descriptive content. This contrast becomes vital when we pay attention to the difference in the number of articles provided by Wikipedia in Spanish and English.

### 4.2.1 Learning Lexico-Syntactic Regularities

Reference [5] computed word association norms directly from unstructured natural language text. They proposed a measure, named *association ratio*, grounded on the idea of mutual information. The *association ratio* ($I_2$) between two words $w_1$ and $w_2$ is defined as:

$$I_2(w_1, w_2) = log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

This ratio compares the probability of observing $w_2$ followed by $w_1$ within a fixed window of $k$ words with the probabilities of observing $w_1$ and $w_2$ independently. This ratio differs from mutual information in the encoded linear precedence, and captures some lexico-syntactic regularities in the target corpus [5].

This ratio is computed by making allowances for a window size of ten, and the probabilities are estimated as described in [5]. Since this ratio becomes unstable when counts are very small, like [5], word pairs with a frequency lower than six were discarded. In addition, pairs consisting solely of stop-words were also filtered out.

Under the underlying assumption that relevant pairs will exhibit a joint probability larger than the product of the probability of finding them by chance. In our work, this word association ratio is extended to triplets as follows:

$$I_3(w_1, w_2, w_3) = log_2 \frac{P(w_1, w_2, w_3)}{P(w_1)P(w_2)P(w_3)}$$

Like [5], we noticed that the larger the ratio is, the more credible results it computes. Conversely, the values become less interesting while the ratio approaches zero. Negative ratios are rare, but possible, and [5] suggests that it indicates a complementary relationship. Simply put, this ratio supplies a efficient way to identify some semantic and lexico-syntactic relations. Table 4 stresses some associations with $w_*$="*politico*" discovered in Spanish.

Incidentally, learning these word association norms raises the issue of orthographical variations in Spanish. The meaning of words can substantially change if they are written with their respective orthographic accents or not. Some good examples are "*corte*" and "*rio*" as well as "*ejercito*". Spanish speakers, however, are likely to omit the orthographical accent when they write on blogs, web documents, or Wikipedia articles. The reason they leave out this accent is that they are normally unnecessary, because the context usually yields enough information to readily disambiguate the correct meaning. For this reason, along with the fact that our tuples represent contextual relations of words, tuples were computed omitting the accent. Another final aspect regarding orthographical variations is misspellings, interchanging "*c*" with "*s*", or "*v*" with "*b*" is very common in Spanish. But unfortunately, this sort of variation is harder to correct, and has an impact on the norms, because the "*new*" word can exist in the Spanish lexicon.

### 4.2.2 Ranking Function

`EsDefWebQA` makes use of algorithm 1 to select answers. However, two aspects of this algorithm must be changed: (a) instead of words, $\phi$ stores pairs and triplets seen in previously selected sentences, and (b) the ranking function is adapted to deal with tuples as follows:

$$R(S_s, \phi) = (1 + \sum_{\forall e_e \in S_s} P(e_e)) * R_I(S_s, \phi)$$

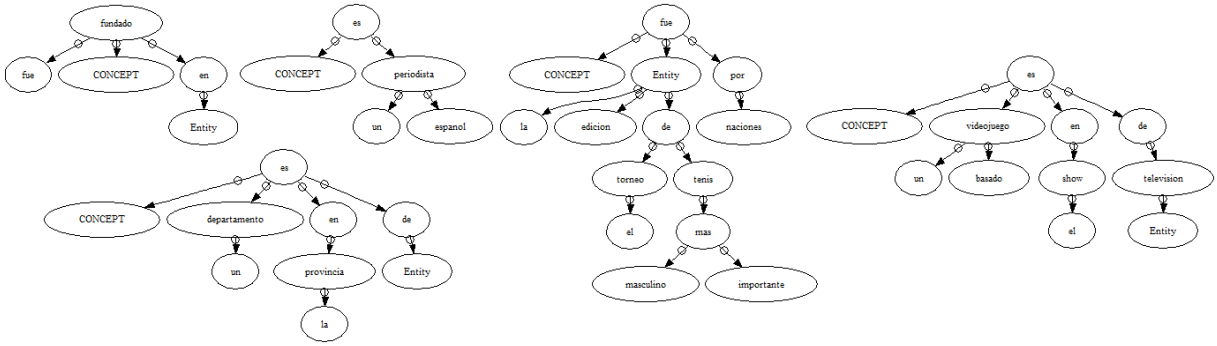


**Figure 1: Some Dependency Trees High in Frequency Extracted from Wikipedia Abstracts.**

This formulae introduces a new factor $R_I(S_s, \phi)$, which is computed according to:

$$R_I(S_s, \phi) = \sum_{\forall \vec{w} \in W_{S_s} - \phi} \bar{I_2'}(\vec{w}) + \bar{I}_3(\vec{w})$$

Where $\vec{w}$ is a tuple taken from $S$. $\bar{I}_2$ and $\bar{I}_3$ correspond to the normalised word association ratios $I_2$ and $I_3$, respectively. `EsDefWebQA` normalises these ratios by dividing by the ratio corresponding to the highest pair and triplet that match sentences in $S$. Then, $\bar{I_2'}(\vec{w})$ is calculated as follows:

$$\bar{I_2'}(\vec{w}) = \begin{cases} \bar{I}_2(\vec{w}) & \text{if } \bar{I}_2(\vec{w}) \neq 0 \\ \bar{H}(\vec{w}) & \text{otherwise} \end{cases}$$

$H$ is an histogram of tuples $\vec{w}$ taken from $S$, where pairs and triplets with a frequency equal to one are removed. This histogram is normalised similarly to the association ratios ($\bar{H}$). The nature of this ranking combines contextual evidence provided by definitions in Wikipedia and some contextual regularities within sentences fetched from the web.

## 4.3 Strategy Three

The previous strategy, on the one hand, combines evidence yielded by candidate sentences with evidence supplied by descriptive sentences across Wikipedia articles. There is still, on the other hand, a big question mark over the word association norms presented in the prior section: extracting pairs and triplets from windows of ten consecutive words starts from the tacit linguistic assumption that lexical dependencies cannot occur between larger spans of words. Intuitively, this problem could be solved by accounting for larger windows, but unfortunately, this would bring out a sharper growth in the amount of tuples. In particular, in the number pairs and triplets corresponding to loosely related words. We deem that this increase will be more prominent than in the amount of tuples of largely related words.

Another valid assumption made by these norms is that a relation between all words within a given window exists. This seems to be utterly reasonable when weakly related tuples are discarded by means of an experimental threshold. However, there are also many significant relationships low in frequency that would be discarded along with these spurious tuples. This is a burning issue when dealing with a training corpus limited in size, because many relevant tuples will obtain a low frequency, and hence look irrelevant.

### 4.3.1 *Learning Lexico-Syntactic Regularities*

In order to surmount these difficulties, a dependency parser is used as an oracle[2] that provides the lexical dependencies in a given descriptive sentence (see [9] for other uses of dependancy parsing for discovering definitions in English). This dependency parser assists in removing the window size and lowering the experimental threshold from six to two. The word association norms are hence computed as pairs and triplets of consecutive words in the dependency paths. Some illustrative examples taken from the dependency trees depicted in figure 1 are:

es→departamento→un
fundado→en→Entity
por→naciones

It is worth noting that dependency paths encode grammatical information about word orderings, however, contrary to the tuples in `Strategy II`, these orderings are not necessarily linear. Since only specific links are taken into account now, the number of tuples declines with respect to the model in section 4.2.1. Table 5 shows this decrease:

| | `Strategy Two` | `Strategy Three` | % |
|---|---|---|---|
| Pairs | 719,510 | 243,286 | 33,81 |
| Triplets | 1,161,743 | 215,119 | 18,52 |

**Table 5: Difference in the Number of Tuples.**

### 4.3.2 *Ranking Function*

This method utilises the same ranking function as the second strategy (see section 4.2.2). However, the word association norms need to be redefined in order to account for tuples taken from the dependency paths:

$$I_2^*(w_1, w_2) = log_2 \frac{P_{link}(w_1, w_2)}{P_g(w_1) P_d(w_2)}$$

Where $P_g(w_1)$ and $P_d(w_2)$ are the probabilities that the word $w_1$ and $w_2$ are independently the head and the dependent, respectively. In addition, $P_{link}$ is the probability of finding the word $w_1$ as the head of $w_2$. Homologous with the second strategy, the number of links in the corpus is

[2]We use FreeLing 2.1 as a dependency parser for Spanish.

| | Strategy One | Strategy Two | Strategy Three |
|---|---|---|---|
| $\mathcal{F}_1$ | 0.40 ± 0.24 | 0.46 ± 0.22 | 0.39 ± 0.19 |
| $\mathcal{F}_2$ | 0.44 ± 0.25 | 0.48 ± 0.22 | 0.42 ± 0.20 |
| $\mathcal{F}_3$ | 0.46 ± 0.26 | **0.49** ± 0.23 | 0.44 ± 0.22 |
| $\mathcal{F}_4$ | 0.47 ± 0.27 | 0.50 ± 0.24 | 0.45 ± 0.22 |
| $\mathcal{F}_5$ | 0.47 ± 0.27 | 0.50 ± 0.24 | 0.46 ± 0.23 |
| Precision | 0.43 ± 0.31 | **0.52** ± 0.34 | 0.41 ± 0.27 |
| Recall | 0.48 ± 0.28 | **0.51** ± 0.25 | 0.47 ± 0.24 |

**Table 6: CLEF 2007 Results ($\mathcal{F}_\beta$ score).**

| | Strategy One | Strategy Two | Strategy Three |
|---|---|---|---|
| $\mathcal{F}_1$ | 0.37 ± 0.18 | 0.32 ± 0.20 | 0.38 ± 0.29 |
| $\mathcal{F}_2$ | 0.43 ± 0.20 | 0.33 ± 0.18 | 0.40 ± 0.22 |
| $\mathcal{F}_3$ | **0.47** ± 0.22 | 0.34 ± 0.18 | 0.41 ± 0.22 |
| $\mathcal{F}_4$ | 0.48 ± 0.23 | 0.35 ± 0.18 | 0.42 ± 0.22 |
| $\mathcal{F}_5$ | 0.50 ± 0.24 | 0.35 ± 0.18 | 0.43 ± 0.22 |
| Precision | **0.55** ± 0.23 | 0.35 ± 0.18 | 0.44 ± 0.23 |
| Recall | 0.37 ± 0.21 | **0.40** ± 0.33 | 0.38 ± 0.29 |

**Table 7: CLEF 2008 Results ($\mathcal{F}_\beta$ score).**

interpreted as the corpus size, when computing the probabilities. Analogously, $I_3^*(w_1, w_2, w_3)$ is defined as:

$$I_3^*(w_1, w_2, w_3) = log_2 \frac{2 * P_{link}(w_1, w_2, w_3)}{P_g(w_1)(P_g(w_2) + P_d(w_2))P_g(w_3)}$$

In a triple, the middle node serves as the dependent and the head of another node. For this reason, both probabilities were averaged. Eventually, when `EsDefWebQA` makes use of this strategy for answering a definition question, it obtains the dependency trees corresponding to sentences in $S$ by means of a parser, extracts the tuples and ranks the sentences according to the equations in section 4.2.2, but making allowances for these new association norms. The extraction algorithm is identical to the second strategy.

## 5. EXPERIMENTS

In order to compare the strategies presented in this work, they were assessed by means of 53 definition questions corresponding to the CLEF 2007-2008 Spanish Question Answering tracks. Since we could not account for the official records, we took into consideration the 19 and 34 questions corresponding to the queries recognised as definitions by the best (INAOE) team in CLEF 2008 and 2007, respectively. Even though we make allowances for CLEF datasets, a TREC-style evaluation is performed. Consequently, for each question, the retrieved snippets were manually inspected (section 3) in order to create a gold standard, like [18]. As evaluation measures, recall, precision and F-score were utilised as presented by [16], and considering nuggets in the gold standard as equally weighed.

Tables 6 and 7 stress the outcomes obtained for each question set. In both cases, `Strategy Two` finished with the highest recall. This means tuples extracted from Wikipedia abstracts contributed to identifying additional descriptive information low in frequency. However, it is crystal clear that this enhancement was modest, but it is nevertheless mildly encouraging. The results are motivating, due to the next two reasons: (a) the number of descriptive sentences utilised for learning tuples is small, and (b) the frequent use of both genders (masculine and feminine) adversely affects our learning models. In English, most nouns have only one neuter form: "*singer*", "*president*" and "*writer*", while few nouns still bear the gender (e. g. "*congressman/congresswoman*"). In Spanish, conversely, most nouns usually carry the gender: "*presidente/presidenta*" and "*escritor/escritora*", whereas few are neuter (e. g. "*cantante*"). This difference in noun forms is vital when having few training examples, because adjectives must agree with the number and the gender of the noun:

```
... intelectual y escritora francesa autora de ...
... filosofo y escritor frances ...
... primera mujer elegida ...
... el primer hombre japones en ...
```

To a limited extent, this problem can be lessened by means of a morphological analyser, such as FreeLing. However, it is worth remarking that Freeling does not provide a mapping to a "*standard*" form for all words (the reader can check this by trying the given examples). In light of this observation, we reasonably deem that boosting the performance will demand considerable efforts. These efforts will go into deeper linguistic processing, and at the same time, collecting a larger set of descriptive sentences. This is contrary to English, where Wikipedia supplies a considerably larger collection and only one gender is predominately used.

Results obtained by `Strategy Three` do not reflect a definite improvement in terms of recall. The reason for this is two-fold: (a) relevant tuples were discarded, when reducing the models, and (b) the dependency paths computed from the candidate set of sentences did not match the paths in the models. In order to corroborate this conclusion, we inspected the average number of matching tuples between the second and third strategies: 124 pairs and 57 triplets for `Strategy Two`, while 20 pairs and 2 triplets for `Strategy Three`. We consider two reasons for this mismatch. Firstly, errors in the output of the parser. Sentences taken from Wikipedia are much more well-formed than -occasionally truncated- phrases within web snippets. Secondly, longer dependency paths might be needed to model the lexical relationships necessary to characterise definitions. However, dealing with these two issues would bring about a significant increase in the retrieval and processing times.

Figure 2 plots the recall versus the number of candidate sentences. This graph shows a satisfactory outcome. Specifically, the performance of `Strategy Two` and `Strategy Three` seems to have an experimental lower "*bound*", when more than thirty candidate sentences exist. Of course, the larger the amount of candidate sentences, the higher the probability of matching our models. However, the interesting aspect here is that this outcome provides a way to automatically determine when it is more reliable to utilise these strategies. At any rate, this stresses the relevance of a massive redundancy in the web collection. Certainly, both strategies suffer from the same drawbacks. But, there is still one extra aspect that it is worth duly pointing out here. Obtaining the dependency trees of the candidate sentences $S$ requires extra computation time when processing the user request.

More interesting conclusions come into light when the results obtained by the first strategy are analysed. `Strategy`

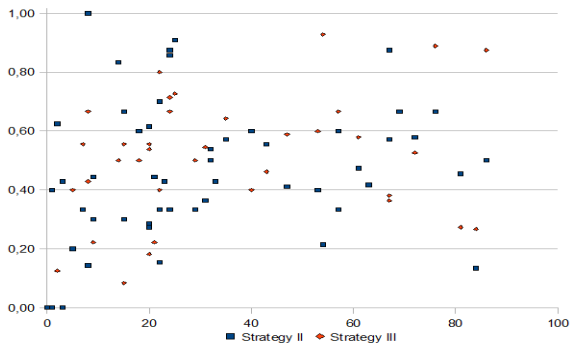


**Figure 2: Recall vs. $|S|$ (`Strategy Two` & `Three`).**

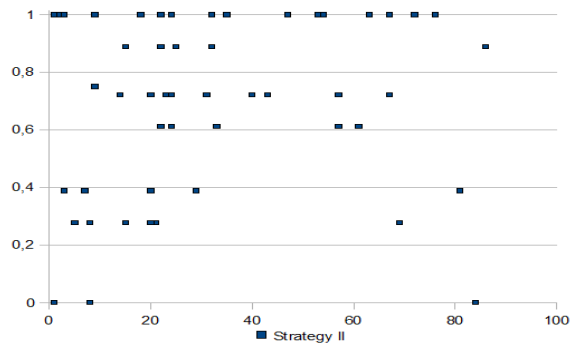


**Figure 3: MAP-3 vs. $|S|$ (`Strategy Two`).**

`One` is based largely on the information fetched from the web, and these results thus show that many definitions can be readily distinguished exclusively from redundancy provided by the web. This exactly means, in some cases, there is no need for an external specific resource or deep linguistic processing to find out information that can assist in discovering definitions (in Spanish) on the web. In essence, our outcomes show that some definitions can be recognised by taking advantage of contextual redundancy and patterns that often convey definitions, even though this contextual redundancy is far from being large-scale.

With regards to precision, results markedly varied from CLEF 2008 to CLEF 2007 and they are thus not conclusive. In order to draw interesting conclusions concerning precision, the *Mean Average Precision* (MAP) of the top ranked and the top three ranked sentences were computed (accounting for "*Precision at one and at three*", respectively). Table 8 highlights these results.

| | `Strategy One` | `Strategy Two` | `Strategy Three` |
|---|---|---|---|
| MAP-1 | 0.62 | 0.69 | 0.65 |
| MAP-3 | 0.58 | 0.66 | 0.62 |

**Table 8: *Mean Average Precision* (MAP).**

The obtained MAP scores show that using our tuples effectively contributes to improving the ranking of the sentences. Essentially, they help to bias the ranking in favour of descriptive sentences that have some lexico-syntactic similarities to sentences in Wikipedia abstracts. A positive aspect of this enhancement in ranking is that our methods are aimed at selecting sentences that yield the more novel and representative content. That is, these three selected sentences are very likely to convey different information, or in the worst case, different paraphrases of the same underlying ideas. This difference can also include several senses. The achieved results hold a promise, because of the small number of training sentences.

One important facet of definition QAS, in particular to search engines, is the MAP for the top ranked sentence. In this aspect, `Strategy Two` outperformed the first and second strategies, ranking a valid definition on the top in 69% of the cases (see table 9 for some examples). Figure 3 plots the *Mean Average Precision* obtained by this strategy versus the number of candidate sentences. This graph shows that this precision tends to be greater than 0.6, when $S$ is greater than thirty. Simply put, in these cases, at least two out of the top three ranked sentences were genuine definitions.

With regards to the $\mathcal{F}_\beta$ score, `Strategy Two` finished with better results than the first and third strategy. In order to use another ranking strategy as reference, the *Centroid Vector* was implemented [6]. This centroid vector was learnt solely from the fetched snippets instead of from external specific resources. The reason for this is three-fold: (a) the limited availability of these external resources for Spanish, (b) other approaches, like [4], have also learnt this vector from web snippets, and more important (c) our goal is determining definitions from their context, without the assistance of an oracle of descriptive information about each particular *definiendum*. Sentences were thereafter ranked according to this vector by means of their cosine similarity. Algorithm 1 was used accordingly, and word included in previously selected sentences were not used for measuring the similarity of the remaining candidate sentences. This ranking strategy finished with $\mathcal{F}_3$ scores 0.18 $\pm$ 0.21 (Precision 0.14 $\pm$ 0.25; Recall 0.21 $\pm$ 0.22) and 0.26 $\pm$ 0.22 (Precision 0.13 $\pm$ 0.16; Recall 0.33 $\pm$ 0.28) for the CLEF 2008 and 2007, respectively.

## 6. CONCLUSIONS

This paper presented several strategies that our system has used for extracting answers to definition questions from the web. Results shows the significant impact of the redundancy of information across the Spanish web.

As current research, the goal is obtaining annotated examples that can be used for discriminant learning. However, there is still no good strategy to obtain good negative examples without manual inspection.

## Acknowledgements

We thank the INAOE team for supplying the question sets utilised in the experiments presented in this work.

This work was partially supported by a research grant from the German Federal Ministry of Education, Science, Research and Technology (BMBF) to the DFKI project `HyLaP` (FKZ: 01 IW F02) and the EC- funded project QALL-ME - FP6 IST-033860 (http://qallme.fbk.eu).

Our definition question answering system can be tested at: `http://experimental-quetal.dfki.de/`, by entering, for instance: "*define:Odessa language:es*".

◇ *Le Corbusier* fue uno de los miembros fundadores del Congreso Internacional de Arquitectura Moderna e hizo famoso el llamado estilo arquitectonico internacional. (http://copypeist.com/2005/08/28/33/)
◇ Gustave *Flaubert* nacio el 12 de diciembre de 1821, en Ruan, Normandia, y murio el 8 de mayo de 1880, en Croisset. (http://es.answers.yahoo.com/question/index?qid=20090105120718AABaKHd)
◇ La *revolucion de los claveles* es el nombre dado al levantamiento militar del 25 de abril de 1974 que provoco la caida en Portugal de la dictadura salazarista que dominaba ..(http://www.estrelladigital.es/ED/diario/51162.asp)
◇ *Marco Pantani* nacio en Cesena, Italia, el 13 de enero de 1970 y debuto como profesional en el Gran Premio de Camaiore para fichar con el equipo Carrera en el que milito desde...
(http://www.esmas.com/deportes/otrosdeportes/343766.html)
◇ La *tarantela* es un baile popular del sur de Italia y, por lo tanto, posiblemente de las regiones italianas de Apulia, Basilicata, Calabria, Molise, Campania o Sicilia.
(http://video.aol.com/video-detail/jascha-heifetz-scherzo-tarantella/1919925688/?icid=VIDURVHOV07)
◇ *INTASAT* es el primer satelite artificial cientifico español. (http://valija-viaje.boonic.com/)
◇ En la *escala de Mohs*, que indica la dureza de los materiales de 1 a 10, el zafiro ocupa la novena posicion por dureza (el diamante tiene 10). (http://www.sobrerelojes.com/TECNICA/relojes-elcristal.htm)
◇ *Leica* es una casa alemana dedicada a la fabricacion de instrumentos opticos de precision.
(http://es.wordpress.com/tag/leica/)
◇ *Odessa* es la tercera ciudad mas grande de Ucrania despues de Kiev y Kharkov, un importante industrial, cultural, cientifico y recurrir centro en el norte de la region del Mar.
(http://www.articleset.com/Recorrido-y-ocio_articles-277_es.htm)
◇ La *vexilologia* es la ciencia que se encarga del estudio de las banderas en todas sus variantes: guiones, estandartes, banderines, vexiloides, etc.
(http://vial.jean.free.fr/new_npi/revues_npi/17_2000/npi_1700/17_spai_vexillo.htm)
◇ Los *pellets* son un nuevo tipo de combustible fabricado de una forma similar a las briquetas de madera, de los desechos de la madera y por medio del prensado... (http://www.atmos.cz/spanish/paliva-energie)
◇ *Rafael Azcona*, que fallecio el pasado 25 de marzo a los 81 años de edad a causa de un cancer de pulmon, es uno de los guionistas mas relevantes en la historia del cine.
(http://actualidad.terra.es/cultura/articulo/se_rafael_azcona_2372591.htm)

**Table 9: Some Top Ranked Sentences (Strategy Two).**

# 7. REFERENCES


[1] A. Juárez-Gonzalez, A. Téllez-Valero, C. Denicia-Carral, M. Montes–y–Gómez and L. Villaseñor–Pineda. INAOE at CLEF 2006: Experiments in Spanish Question Answering. In *Working Notes for the CLEF 2006 Workshop*, 2006.

[2] A. Téllez, A. Juárez, G. Hernández, C. Denicia, E. Villatoro, M. Montes L. Villaseñor. INAOE's Participation at QA@CLEF 2007. In *Working Notes for the CLEF 2007 Workshop*, 2007.

[3] B. Magnini, D. Giampiccolo, P. Forner, C. Ayache, V. Jijkoun, P. Osenova, A. Peñas, P. Rocha, B. Sacaleanu and R. Sutcliffe. Overview of the CLEF 2006 Multilingual Question Answer ing Track. In *Working Notes for the CLEF 2006 Workshop*, 2006.

[4] Y. Chen, M. Zhon, and S. Wang. Reranking Answers for Definitional QA Using Language Modeling. In *Coling/ACL-2006*, pages 1081–1088, 2006.

[5] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. In *Computational Linguistics, Vol. 16, No. 1*, pages 22–29, 1990.

[6] T. Cui, M. Kan, and J. Xiao. A comparative study on sentence retrieval for definitional question answering. In *SIGIR Workshop on Information Retrieval for Question Answering (IR4QA)*, pages 383–390, 2004.

[7] D. Giampiccolo, A. Peas, C. Ayache, D. Cristea, P. Forner, V. Jijkoun, P. Osenova, P. Rocha, B. Sacaleanu and R. Sutcliffe. Working Notes for the CLEF 2007 Workshop, 19-21 September, Budapest, Hungary. In *Working Notes for the CLEF 2007 Workshop*, 2007.

[8] C. Denicia-Carral, M. M. y Gómez, L. V. Pineda, and R. Hernández. A Text Mining Approach for Definition Question Answering. In *FinTAL*, pages 76–86, 2006.

[9] A. Figueroa and J. Atkinson. Using dependency paths for answering definition questions on the web. In *Proceedings of the Fifth International Conference on Web Information Systems and Technologies*, 2009.

[10] A. Figueroa and G. Neumann. A Multilingual Framework for Searching Definitions on Web Snippets. In *KI 2007: Advances in Artificial Intelligence, 30th Annual German Conference on AI*, pages 144–159, 2007.

[11] Z. Harris. Distributional structure. In *Distributional structure. Word, 10(23)*, pages 146–162, 1954.

[12] W. Hildebrandt, B. Katz, and J. Lin. Answering Definition Questions Using Multiple Knowledge Sources. In *HLT-NAACL*, pages 49–56, 2004.

[13] H. Joho and M. Sanderson. Retrieving Descriptive Phrases from Large Amounts of Free Text. In *9th ACM conference on Information and Knowledge Management*, pages 180–186, 2000.

[14] H. Joho and M. Sanderson. Large Scale Testing of a Descriptive Phrase Finder. In *1st Human Language Technology Conference*, pages 219–221, 2001.

[15] W. Kintsch. Predication. *Cognitive Science*, 25:173–202, 1998.

[16] J. Lin and D. Demner-Fushman. Will pyramids built of nuggets topple over? In *Proceedings of the main conference on HTL/NAACL*, pages 383–390, 2006.

[17] M. M. Soubbotin. Patterns of Potential Answer Expressions as Clues to the Right Answers. In *Proceedings of the TREC-10 Conference*, 2001.

[18] E. M. Voorhees. Evaluating Answers to Definition Questions. In *HLT-NAACL*, pages 109–111, 2003.

[19] J. Xu, Y. Cao, H. Li, and M. Zhao. Ranking Definitions with Supervised Learning Methods. In *WWW2005*, pages 811–819, 2005.