A multilingual framework for searching definitions on web snippets

Alejandro Figueroa and Günter Neumann

Deutsches Forschungszentrum für Künstliche Intelligenz - DFKI,
Stuhlsatzenhausweg 3, D - 66123, Saarbrücken, Germany
{figueroa|neumann}@dfki.de

Abstract. This work[1] presents Mdef-WQA, a system that searches for answers to definition questions in several languages on web snippets. For this purpose, Mdef-WQA biases the search engine in favour of some syntactic structures that often convey definitions. Once descriptive sentences are identified, Mdef-WQA clusters them by *potential senses* and presents the most relevant phrases of each *potential sense* to the user. The approach was assessed with TREC and CLEF data. As a result, Mdef-WQA was able to extract descriptive information for all definition questions in the TREC 2001 and 2003 data-sets.

1 Introduction

In recent years, search engines have considerably improved their power of indexing in response to the constantly increasing number of documents on the Internet and the growing need of users for smarter ways of searching and presenting the information. Nowadays, one pressing need is to find definitions of concepts. High-performance search engines, such as Google, provide hence a feature which helps users to retrieve definitions from specialised online resources like WordNet and Wikipedia. Google is additionally urged to supply an interface of Wikipedia in other languages, in order to satisfy users all around the world.

Google relies upon the coverage and the high cachet of these specialised resources, especially upon the fact that the first sentence they provide is extremely likely to yield a definition. Unfortunately, this coverage tremendously varies over languages. For instance Wikipedia contains more than 1700000 articles in English whereas about 220000 in Spanish. Further, Google does not make allowances for the redundancy on the responses (i. e. "*George Bush*" in English). Furthermore, Google provides undesirable definitions for some well-known concepts, for example "*George Bush*" in German. Moreover, Google does not present to the user definitions grouped by their respective senses (i. e. "*Tesla*").

During the last years, the problem of finding definitions for a specific concept (the *definiendum*) has been addressed by Question Answering Systems (QASs)

[1] The work presented here 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.

in the context of the Text REtrieval Conference (TREC) and the Cross Language Evaluation Forum (CLEF). In TREC, QASs answer definition questions in English, such as "*What is a quasar?*", by extracting as much as possible non-redundant descriptive information ('nuggets') about the *definiendum* from the ACQUAINT corpus.

In order to discover definition utterances, definition QASs usually align sentences with surface patterns in the target corpus at the word and/or the part-of the speech level [5]. Hence, the probability of matching sentences increases as long as the size of the target collection grows, and accordingly, the performance substantially improves [6]. Along with surface patterns, definition QASs take advantage of wrappers around online resources, WordNet glossaries and web snippets [1]. In addition, QASs, like Google, have also shown that definition web-sites are a fertile source of descriptive information in English, in particular, providing answers to 42 out of 50 TREC–2003 questions [1]. However, web snippets have not yet proven to be a valuable source of descriptive phrases so far [1].

Full documents have also been used for extracting definitions. For example, in [10], 250-characters long windows that convey a definition are obtained from the top 50 documents fetched by an IR engine. The windows were then ranked by a *Support Vector Machine*, which was trained using previously tagged windows according to the criteria of [6], and some automatically acquired phrasal attributes. This system obtained one acceptable definition within the top-five ranked windows for 116 out of 160 TREC–2000 questions and 116 out of 137 TREC–2001 questions.

However, TREC focuses its attention solely on English, whereas CLEF aims essentially at European Languages. In the context of CLEF, surface patterns have also shown to be useful for recognising descriptive sentences in other languages. For instance, the best system in CLEF–2005 answered 40 out of the 50 definition questions in the Spanish track by means of surface patterns [13, 11].

QASs normally tackle redundancy by: (a) randomly removing one sentence from every pair that shared more than 60% of their terms [5], or (b) filtering out candidate sentences by ensuring that their cosine similarity to all previously selected utterances is below a threshold. It is also worth to remark that definition QASs have not yet made effort to deal with the disambiguation of the different senses of the *definiendum*.

Our contribution

Unlike current definition QASs or search engines, we propose a QAS (named `Mdef-WQA`) that extracts descriptive phrases *directly* from web snippets by rewriting the prompted query in such a way that the probability of aligning surface patterns with web snippets increases (i. e. snippets from specialised definition web-sites like Wikipedia). Since `Mdef-WQA` bases its search on the efficiency of surface patterns and its coverage on the entire web, we show that the framework of `Mdef-WQA` is applicable to *several languages*, in particular English and

Spanish. Moreover, we present a novel approach to cluster descriptive utterances according to *potential senses*, which are used to provide a partition of the most relevant and diverse utterances to the user. `Mdef-WQA` was evaluated in detail using the TREC and CLEF data-sets. The results show that `Mdef-WQA` is promising for answering definition questions in several languages directly from web snippets. In particular, `Mdef-WQA` found out descriptive information for all definition questions in the TREC 2001 and 2003 data sets.

2 Mining the web for definitions

Like [10], `Mdef-WQA` receives the *definiendum* δ as input, assuming that it is previously identified by an external query analysis module or entered by the user. Analogously, `Mdef-WQA` receives the language ζ of the original query Q, because it cannot be inferred directly from δ, especially for proper names (i. e. "*John Kennedy*"). `Mdef-WQA` proceeds then as follows:

1. `Mdef-WQA` uses δ and ζ for **rewriting** Q according to a set Π^{ζ} of pre-defined surface patterns for ζ. These generated queries are then submitted to the search engine. This **rewriting** boosts the retrieval of descriptive utterances by biasing the search engine in favor of sentences that match Π^{ζ}. Hence, `Mdef-WQA` **avoids** the implementation of **specialised wrappers** and **downloading full documents**, contrary to the trend of current definition QASs.
2. `Mdef-WQA` aligns these patterns with sentences in fetched snippets. Due to its complex internal structure [12], δ might match the *definiendum* δ' only partially within the retrieved descriptive utterances. `Mdef-WQA` recognises δ by means of relaxed pattern matching based on the *Jaccard Measure*. The motivation for using this relaxed matching strategy is that it provides `Mdef-WQA` with a higher degree of language independence compared to current definition QAS. In particular, we avoid the specification of additional word addition/ordering rules [12] or the integration of more sophisticated linguistic processing such as chunking [5].
3. `Mdef-WQA` groups sentences by *potential senses*, which are discovered by **observing the partitions generated by the closest neighbours** of δ in the reliable semantic space supplied by Latent Semantic Analysis (LSA). LSA supplies of language independent framework for drawing semantic inferences.
4. `Mdef-WQA` takes advantage of a variation of Multi-Document Maximal Marginal Relevance [4] for reducing redundancy and maximising diversity in selected utterances. This guarantees a fast summarisation framework which only makes use of a language–specific stop-list.

2.1 Obtaining descriptive sentences

In recent years, surface patterns for English have proven to be useful for distinguishing definition utterances in natural language texts [12, 10, 5–7]. These surface patterns provide syntactic structures that are properly aligned with sentences in order to detect descriptive utterances. The syntactic structures are,

Table 1. Surface Patterns for English (Π^{en}).

π_1^{en}: $\delta^{'}$ [is|are|has been|have been|was|were] [a|the|an] $\eta^{'}$
e.g.,"***Noam Chomsky*** *is a writer and critic...*"
π_2^{en}: $[\delta^{'}|\eta^{'}]$, [a|an|the] $[\eta^{'}|\delta^{'}]$ [,|.]
e.g.,"***The new iPoD****, an MP3-Player,...* "
π_3^{en}: $\delta^{'}$ [become|became|becomes] $\eta^{'}$
e.g.,"***In 1957, Althea Gibson*** *became the...*"
π_4^{en}: $\delta^{'}$ [which|that|who] $\eta^{'}$
e.g.,"***Joe Satriani*** who was inspired to play..."
π_5^{en}: $\delta^{'}$ [was born] $\eta^{'}$
e.g.,"***Alger Hiss*** was born in 1904 in USA..."
π_6^{en}: $[\delta^{'}|\eta^{'}]$, or $[\eta^{'}|\delta^{'}]$
e.g.,"***Sting****, or Gordon Matthew Sumner,...*"
π_7^{en}: $[\delta^{'}|\eta^{'}][|,][|$also|is|are] [called|named|nicknamed|known as] $[\eta^{'}|\delta^{'}]$
e.g.,"***Eric Clapton****, nicknamed 'Slowhand'...*"
π_8^{en}: $[\delta^{'}|\eta^{'}]$ $([\eta^{'}|\delta^{'}])$
e.g.,"***The United Nations*** *(UN)..*"

more precisely, based largely upon punctuation and words that often convey definitions. Simply put, these syntactic structures make available the way to identify the *definiendum* $\delta^{'}$ and its definition nugget $\eta^{'}$ within utterances.

`Mdef-WQA` takes advantage of these syntactic structures not only for distinguishing definitions, but also for biasing the search engine in favor of web snippets that convey definitions. Table 1 shows surface patterns that we found to be particularly useful for this purpose. From this manually specified set of patterns, `Mdef-WQA` automatically generates the following set of ten different queries used by the search engine. The first submission q_1 corresponds to "δ", and the next four queries aims at π_1^{en}:

q_2:"δ is a " $\vee$ "δ was a " $\vee$ "δ were a " $\vee$ "δ are a "
q_3:"δ is an " $\vee$ "δ was an " $\vee$ "δ were an " $\vee$ "δ are an "
q_4:"δ is the " $\vee$ "δ was the " $\vee$ "δ were the " $\vee$ "δ are the "
q_5:"δ has been a " $\vee$ "δ has been an " $\vee$ "δ has been the " $\vee$ "δ have been a " $\vee$ "δ have been an " $\vee$ "δ have been the "

π_1^{en} is split into four queries, because it retrieves many descriptive utterances. The next query q_6 attempts to discover snippets that match π_2^{en} or π_6^{en}:

q_6:"δ, a " $\vee$ "δ, an " $\vee$ "δ, the " $\vee$ "δ, or "

The reason to merge these two patterns into one query is two-fold: (a) π_6^{en} has a low occurrence within web snippets (see also [7]), and (b) π_6^{en} often yields a synonym of δ (i. e. "*myopia, or nearsightedness*"). Alternative names of persons, organisations or abbreviations are seldom expressed in this way, but are likely to match the other clauses within q_6. Consequently, the combination of both

patterns helps `Mdef-WQA` to reduce the number of search calls. The queries q_7, q_8 and q_9 aim at π_7^{en}, π_3^{en} and π_4^{en} respectively as follows:

q_7:(“δ” $\vee$ “δ also ” $\vee$ “δ is ” $\vee$ “δ are ”) $\wedge$ (called $\vee$ nicknamed $\vee$ “known as”)
q_8:“δ became ” $\vee$ “δ become ” $\vee$ “δ becomes ”
q_9:“δ which ” $\vee$ “δ that ” $\vee$ “δ who ”

Finally, q_{10}:“δ was born ” $\vee$ “(δ)” attempts to fetch snippets that match π_5^{en} and π_8^{en}. Similarly to q_6, `Mdef-WQA` merges both patterns into one query on the ground that π_5^{en} deals with δ regarding persons and π_8^{en} focuses basically on acronyms [7]. Hence, `Mdef-WQA` avoids an unproductive retrieval without diminishing the number of fetched descriptive sentences.

Surface patterns for English have been studied widely, especially in TREC, whereas patterns for other languages have been systematically explored only in the context of the CLEF campaigns. Until 2005, CLEF focused exclusively on definition questions aiming at abbreviations and the position of persons [9, 13]. These surface patterns are therefore specialised for recognising this specific sort of descriptive information. Systems in TREC are encouraged in extracting as much as possible useful descriptive information about δ [5]. Thus, these surface patterns provide a wider coverage than patterns known for other languages.

For the particular case of surface patterns for Spanish, two additional issues complicates the identification of descriptive information from the web. Firstly, the patterns are based largely upon punctuation signs [11] and closed class words [3], which are usually ignored by some search engines. Secondly, these punctuation signs and closed class words tend to be separated by a large span of text, which usually contains δ' and/or its respective definition η'. Therefore, supplying syntactic structures seems to be unsuitable for rewriting the query. An illustrative example is the pattern “*El* η'*,* δ'*, se*”, which matches sentences such as “*El presidente de España, Jose Luis Zapatero, se...*”. The snippets obtained by the respective query rewriting “*El*” $\wedge$ “*, δ, se*” are unlikely to yield definitions, and additionally, portions of the large span of text between δ and the closed class word “*El*” can be replaced with an intentional break (often denoted by ...) by the search engine.

All things considered, `Mdef-WQA` seeks to explore whether the translation of surface patterns from English to Spanish provide a wider coverage, and whether they are more efficient for retrieving sentences that convey definitions from the web. Table 2 shows the respective translations of the first five patterns π_p^{en} to Spanish. The translations of π_6^{en} and π_7^{en} as well as some translations of π_3^{en} were not taken into account, because we found them to be unlikely to occur within web snippets. π_8^{en}, which actually does not need any translation, was deliberately omitted for two reasons: it is commonly used by systems in CLEF for resolving abbreviations [11], and one of the motivations behind our research is measuring the contribution of the translated patterns.

From table 2 it can also be observed that pattern π_1^{es} generates 60 cues (e.g., “*es la*”, “*es lo*”, “*son una*”), in contrast to its homologous π_1^{en}, which brings about 18 cues. This substantial increase is due to the fact that Spanish is morphologically richer than English causing a decisive impact on the form and number of queries that `Mdef-WQA` must submit to the web. `Mdef-WQA` necessarily needs to regulate the trade-off between recall and retrieval time. Thus, it is unfeasible to send each cue individually to the web or to follow a criteria similar to the one used for designing the queries for English, because of the number of cues and the fact that they do not present any usefull disjunction. Consequently, the next three key aspects were considered for

Table 2. Surface Patterns for Spanish (Π^{es}).

π_1^{es}: $\delta^{'}$ [es|son|fueron|fue|ha sido|han sido] [la|lo|el|un|una|uno|unos|unas|las|los] $\eta^{'}$
e.g.,"***Jose Luis Zapatero*** *es el relevo de Felipe Gonzalez para los socialistas.*"
π_2^{es}: $\delta^{'}$ [,|;] [un|una|uno|la|lo|el|los|las] $\eta^{'}$ [,|;|.]
e.g.,"***Silvio Rodriguez****, uno de los exponentes de la Nueva Trova cubana,... "*
π_3^{es}: $\delta^{'}$ [ha llegado a ser|llego a ser|se transformo|se ha transformado] $\eta^{'}$
e.g.,"***España*** *se ha transformado en un pais democratico.*"
π_4^{es}: $\delta^{'}$ [,|] [el cual|la cual|los cuales|quien|que] $\eta^{'}$
e.g.,"***Michelle Bachelet*** quien es la primera presidenta de la historia de Chile,..."
π_5^{es}: $\delta^{'}$ [nacio|fue fundado|fue fundada] $\eta^{'}$
e.g.,"***Jose Luis Rodriguez Zapatero*** nacio en Valladolid el 4 de Agosto de 1960."

Table 3. Generated queries for Spanish.

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

designing the queries (table 3): (a) cues that are more likely to retrieve descriptive utterances are distributed in different queries, and some unproductive combinations in π_1^{es} are discarded, (b) cues aiming at different tenses and genders were also spread over different queries; this way `Mdef-WQA` decreases the number of fruitless retrievals, and (c) the number of clauses in a query is limited by the length of queries accepted by search engines.

Once all snippets are fetched `Mdef-WQA` removes all orthographic accents and splits them into sentences by means of intentional breaks and a sentence splitter.[2] Patterns are then applied to discriminate descriptive utterances within retrieved snippets. Since δ does not exactly match $\delta^{'}$, `Mdef-WQA` takes advantages of the *Jaccard Measure* for distinguishing more reliable descriptive sentences. The *Jaccard Measure* J of two terms w_i, w_j is the ratio between the number of different *uni-grams* that they share, and the total number of different *uni-grams*: $J(w_i, w_j) = \frac{|w_i \cap w_j|}{|w_i \cup w_j|}$. Consider for example the *definiendum* δ^*="*John Kennedy*", which might also be expressed as $\delta_1^{'*}$="*John Fitzgerald Kennedy*" or $\delta_2^{'*}$="*Former US President Kennedy*". The values for $J(\delta^*, \delta_1^{'*})$

[2] We are using the one provided by JavaRAP, cf. http://www.comp.nus.edu.sg/∼qiul/NLPTools/JavaRAP.html.

and $J(\delta^*, \delta_2^{'*})$ are $\frac{2}{3}$ and $\frac{1}{5}$ respectively. `Mdef-WQA` filters reliable descriptive utterances by means of a pattern specific threshold, avoiding additional purpose-built hand-crafted rules and ad-hoc linguistic processing. Of course, some sentences containing useful nuggets will be discarded, but these discarded nuggets can also be found in other retrieved phrases, e.g., "*Former US President Kennedy*" in "*John Fitzgerald Kennedy was a former US President.*". In short, `Mdef-WQA` trusts implicitly in the redundancy of the web for discovering several paraphrases.

2.2 Potential Senses Identification

There are many-to-many mappings between names and their concepts. On the one hand, the same name or word can refer to several meanings or entities. On the other hand, different names can indicate the same meaning or entity. To illustrate this, consider the next set S of recognised descriptive utterances:

1. John Kennedy was the 35th President of the United States.
2. John F. Kennedy was the most anti-communist US President.
3. John Kennedy was a Congregational minister born in Scotland

In these sentences, "*US President John Fitzgerald Kennedy*" is referred to as "*John Kennedy*" and "*John F. Kennedy*", while "*John Kennedy*" indicates also a Scottish Congregational minister. In the scope of this work, a *sense* is one meaning of a word or one possible reference to a real-world entity.

`Mdef-WQA` disambiguates senses of δ by observing the correlation of its neighbours in the reliable semantic space provided by LSA. This semantic space is constructed from the term-sentence matrix M, which considers δ as a *pseudo-sentence* which is weighted according to the traditional *tf-idf*. `Mdef-WQA` builds the dictionary of terms W from normalised elements in S, which consists of uppercasing, removal of html-tags, and the isolation of punctuation signs. `Mdef-WQA` distinguishes then all possible different *n-grams* in S together with their frequencies. The size of W is then reduced by removing *n-grams*, which are substrings of another equally frequent term. This reduction allows the system to speed up the computation of M as $UDV^{'}$ using the *Singular Value Decomposition*. Furthermore, the absence of syntactical information of LSA is slightly reduced by considering strong local syntactic dependencies.

`Mdef-WQA` makes use of $\hat{D}$, the greatest three eigenvalues of D, and the corresponding three vectors $\hat{U}$ and $\hat{V}$ for constructing the semantic space as $R = \hat{U}\hat{D}^2\hat{U}^{'}$. `Mdef-WQA` prefers the dot product above the traditional cosine as a measure of the semantic relatedness $R(w_i, w_j) = \hat{u_i}\hat{D}^2\hat{u_j}^{'}$ $(\hat{u_i}, \hat{u_j} \in \hat{U})$ of two terms $w_i, w_j \in W$. The major reasons are (a) it was observed experimentally that, because of the size of web snippets (texts shorter than 200 words), the cosine draws an unclear distinction of the semantic neighbourhood of δ, bringing about spurious inferences [15], and (b) the length of vectors was found to draw a clearer distinction of the semantic neighbourhood of δ as this biases R in favour of contextual terms, which LSA knows better [2].

In this semantic space, the neighbourhood of a particular word w_i provides its context [2, 8]. Consequently, it determines its right meaning by pruning, for instance, inappropriate senses [8]. Similarly, δ is also a term defined by its neighbourhood in this semantic space. For this reason, `Mdef-WQA` selects a set $\bar{W} \subseteq W$ of the forty highest closely related terms to δ, that is, terms that are likely to define its meaning. However, as a result of the relaxed pattern matching, `Mdef-WQA` must also account for all *n-grams* $\delta^+ \in W$ in δ, because some internal *n-grams* could be more likely to

occur within descriptive utterances (i.e., names or surnames are more frequent than their respective full names). In our working sentences and illustrative variations of δ, "*Kennedy*" has a higher frequency than "*John Kennedy*". `Mdef-WQA` considers therefore the forty highest pairs $\{w_i, R_{max}(\delta, w_i)\}$, where $R_{max}(\delta, w_i) = \max_{\delta^+ \in W} R(\delta^+, w_i)$. `Mdef-WQA` normalises terms in $\bar{W}$ according to:

$$\hat{R}(\delta, w_i) = \frac{R_{max}(\delta, w_i)}{\sum_{\forall w_j \in \bar{W}} R_{max}(\delta, w_j)}$$

Since words that indicate the same sense co-occur, `Mdef-WQA` identifies *potential senses* by finding a set $\bar{W}^\lambda \subseteq \bar{W}$ of words, for which their vectors form an orthonormal basis. In order to discriminate these orthonormal terms, `Mdef-WQA` builds a term-sentence matrix Φ, where a cell $\Phi_{is} = 1$, if the term $w_i \in \bar{W}$ occurs in the descriptive phrase $S_s \in S$, zero otherwise. The degree of correlation amongst words in $\bar{W}$ across S is then given by $\hat{\Phi} = \Phi\Phi'$. For example, for the words in $\bar{W}$: w_1= "*Scotland*", w_2= "*President*" and w_3= "*35th*", the computed values for Φ and $\hat{\Phi}$ are:

$$\Phi = \begin{pmatrix} & S_1 & S_2 & S_3 \\ w_1 & 0 & 0 & 1 \\ w_2 & 1 & 1 & 0 \\ w_3 & 1 & 0 & 0 \end{pmatrix} \quad \hat{\Phi} = \begin{pmatrix} & w_1 & w_2 & w_3 \\ w_1 & 1 & 0 & 0 \\ w_2 & 0 & 2 & 1 \\ w_3 & 0 & 1 & 1 \end{pmatrix}$$

Hence, the number of non-selected words $w_j \in \bar{W} - W^\lambda$ that co-occur with a term $w_i \in \bar{W}$ across S is given by:

$$\gamma(w_i) = \sum_{\forall w_j \in \bar{W} - \bar{W}^\lambda : \hat{\Phi}_{ij} > 0} 1$$

In our working example, $\gamma(w_1) = 1$ and $\gamma(w_2) = \gamma(w_3) = 2$, because "*President*" and "*35th*" co-occur in S_1, and "*Scotland*" does not co-occur with any other element of $\bar{W}$. Then, `Mdef-WQA` adds the w_i to $\bar{W}^\lambda$ that:

$$\max_{w_i \in \bar{W}} \gamma(w_i) \tag{1}$$

subject to:

$$\hat{\Phi}_{ij} = 0, \ \forall w_j \in \bar{W}^\lambda \tag{2}$$

$$\gamma(w_i) > 0 \tag{3}$$

In words, a term w_i signals a new sense, if it does not co-occur at the sentence level with any other already selected term $w_j \in \bar{W}^\lambda$, and it has the highest number of co-occurring non-selected terms $w_j \in \bar{W}$. Incidentally, `Mdef-WQA` breaks ties by randomly selecting a term. In our illustrative example, if w_3 is randomly selected, then $\gamma(w_i)$ is equal to one for the three words in the next cycle. w_1 is then selected, because w_3 was already selected and w_2 co-occurs with w_3 ($\hat{\Phi}_{23} > 0$), and accordingly, $\bar{W}^\lambda$ is {"*Scotland*", "*35th*"}. Words are added to $\bar{W}^\lambda$ until no other term w_i fulfils conditions (2) and (3). Next, sentences are divided into clusters C_λ according to terms in $\bar{W}^\lambda$. Sentences that do not contain any term in $\bar{W}^\lambda$ are collected in a special cluster C_0. For our working example, the clusters are: C_0= $\{S_2\}$, C_1= $\{S_3\}$ and C_2= $\{S_1\}$.

Finally, `Mdef-WQA` attempts to reassign each sentence S_s in C_0 by searching for the strongest correlation between its named entities (NEs) and the NEs of a cluster C_λ:

$$\max_{C_\lambda} \sum_{\forall e \in S_s} freq_{C_\lambda}(e) > 0, \quad \lambda \neq 0$$

where $freq_{C_\lambda}(e)$ is the frequency of NEs e in the cluster C_λ. The assumption here is that the same NEs tend to occur in the same sense. To illustrate this, S_2 is assigned to C_2.

2.3 Redundancy Removal

For each cluster C_λ, `Mdef-WQA` determines incrementally a set Θ_λ of its sentences S_λ to maximise their comparative relevant novelty:

$$\max_{S_s \in S_\lambda - \Theta_\lambda} coverage(S_s) + content(S_s)$$

subject to:

$$coverage(S_s) \geq \psi^* > 0 \tag{4}$$

$$W_{type}(S_s) = 0 \tag{5}$$

The comparative relevant novelty of a sentence S_s is given by the relative coverage and content of its nuggets respecting Θ_λ. Let $N(S_s)$ be the set of normalised nuggets associated with S_s and W_N then the **set** of terms of all normalised nuggets. $W_{N(S_s)}$ is the **set** of words in $N(S_s)$. Coverage is then defined as follows:

$$coverage(S_s) = \sum_{\forall w_i \in W_{N(S_s)} - W_{\Theta_\lambda}} P_i$$

where P_i is defined as the probability of finding a word $w_i \in W_N$, and is arbitrarily set to zero for all stop words. W_{Θ_λ} is the **set** of words occurring in preceding selected sentences Θ_λ.

Coverage aims at measuring how likely are novel terms (not seen in Θ_λ) within $N(S_s)$ to belong to a description. Thus, diverse sentences are preferred over sentences with many redundant words, which are consequently filtered according to an experimental threshold ψ^*. On the other hand, content discriminates the degree, in which $N(S_s)$ conveys definition aspects of δ based upon highly close semantic terms and entities, and is given by:

$$content(S_s) = \sum_{\forall w_i \in \bar{W}} \Phi_{is} \hat{R}(\delta, w_i) + \sum_{\forall e \in N(S_s) - E_\lambda} P_e^\lambda$$

The first sum measures the semantic bonding of terms in the respective nuggets, and the second sum the relevance of novel entities (E_λ is the set of entities in Θ_λ). Each novel entity e is weighed according to its probability P_e^λ of being in the normalised nuggets of C_λ. Incidentally, $W_{type}(S_s)$ is the amount of undesirable symbols in S_s such as pronouns, unclosed brackets or parenthesis, URLs. Consequently, condition 5 bans sentences containing such symbols from Θ_λ. In sum, `Mdef-WQA` ranks sentences according to the order they are inserted into Θ_λ. This means that higher ranked sentences

are more diverse, less redundant, and are likely to contain entities along with terms that describe aspects of δ.

Note further that C_0 is processed last in order to initialise Θ_λ with all sentences selected from previous clusters, so that only sentences with novel pieces of information remain in C_0.

3 Experiments and Results

`Mdef-WQA` was assessed by means of standard question sets.[3] The following data sets were considered for English: (1) TREC 2001, (2) TREC 2003, (3) CLEF 2004, (4) CLEF 2005, and (5) CLEF 2006. For Spanish only (4) and (5) were taken into account. All surface patterns thresholds were set to 0.25, apart from thresholds for patterns π_1^{en}, π_5^{en}, π_1^{es} and π_4^{es}, which were set to 0.33, 0.5, 0.33 and 0.4 respectively. These values were determined after experimentally testing different thresholds from 0.2 to 0.7, and thus manually counting the corresponding number of non-descriptive or spurious selected sentences. The threshold that controls redundancy ψ^* was set to 0.01 for both languages.

Three baselines were designed, one for English (`Baseline EN-I`) and two for Spanish (`Baseline ES-I` and `Baseline ES-II`). Like `Mdef-WQA`, `Baseline EN-I` retrieves 300 hundred snippets by submitting "δ" to the web. The retrieved snippets are split into sentences by means of JavaRAP, interpreting intentional breaks as end of sentences. `Baseline EN-I` also accounts solely for a stricter matching of δ by setting all pattern Π^{en} thresholds to one. A random sentence from a pair that shares more than 60% of their terms is discarded, cf. [5], as well as sentences that are a substring of another sentence. `Baseline ES-I` and `Baseline ES-II` do the same processing as `Baseline EN-I`, but they retrieve 420 snippets. These two baselines also differ from `Baseline EN-I` in the number of terms that two sentences must share to be considered as redundant. They account for a threshold of 90% instead of 60%, because the coverage of web space for Spanish is smaller than English and some relevant nuggets are missed along with the redundant content. The difference between the Spanish baselines is that `Baseline ES-I` aims at Π^{es} whereas `Baseline ES-I` at the patterns in [11].

In general, `Mdef-WQA` outputs short sentences, in particular, output sentences for English are comparative longer than the 100 characters (without considering white spaces) nuggets of [5] and smaller than the 250 characters (considering white spaces) fixed windows of [10]. Given the lengths of the outputs of `Baseline EN/ES-I` and `Mdef-WQA EN/ES` (see table 4), it can be concluded that the increase indicates that `Mdef-WQA` **outputs more complete sentences**, lessening the effects of intentional breaks on web snippets. Due to the acceptable length of descriptive sentences and the fact that many nuggets seems odd without their context [5], `Mdef-WQA` outputs sentences instead of only nuggets.

The degree of redundancy of a sentence S_s was roughly approximated at the word level by looking for a sentence $S_{s'}$ in the same response that shares the maximum number of terms with S_s:

$$redundancy(S_s) = \max_{S_{s'} \neq S_s} \frac{ns(S_s \cap S_s^{'})}{ns(S_s)}$$

[3] Along this section, $\pm$ stands for standard deviation, and CLEF data-sets consider all English translations from all languages.

Table 4. Length of output sentences.

	with white spaces	without white spaces
`Baseline ES-I`	98.11 ± 44.90	81.06 ± 37.69
`Baseline ES-II`	104.98 ± 36.43	85.88 ± 29.87
`Mdef-WQA ES`	135.78 ± 45.21	113.70 ± 37.97
`Baseline EN-I`	118.168 ± 50.20	97.81 ± 41.80
`Mdef-WQA EN`	125.70 ± 44.21	109.74 ± 42.15

where $ns(S_s)$ is the number of words in S_s excluding stop-words. As a result, `Baseline ES-II` generates an output, at least, two times redundant as `Mdef-WQA`, which supplies longer sentences (see table 5). By and large, `Mdef-WQA` **outputs comparative longer and less redundant sentences**.

Table 5. Redundancy overview.

	(1)	(2)	(3)	(4)	(5)
`Baseline ES-I`				0.32 ± 0.16	0.38 ± 0.25
`Baseline ES-II`				0.54 ± 0.24	0.64 ± 0.39
`Mdef-WQA ES`				0.25 ± 0.17	0.25 ± 0.16
`Baseline EN-I`	0.58 ± 0.26	0.61 ± 0.26	0.57 ± 0.25	0.62 ± 0.25	0.53 ± 0.23
`Mdef-WQA EN`	0.47 ± 0.18	0.50 ± 0.20	0.45 ± 0.18	0.45 ± 0.17	0.45 ± 0.19

The coverage of surface patterns for English has been studied widely [5–7], by the same token table 6 shows the number of descriptive sentences in the final output that match each pattern in Π^{es}. Each cell represents the number of matches for the CLEF 2005/2006 corpus respectively. π_1^{es} provides the wider coverage, while π_3^{es} the most limited. Given the marked increase in the number of recognised descriptive utterances in the final output, it can be concluded that our query rewriting strategy strongly biases the search engines not only in favour of **redundant** descriptive sentences, but also in favour of **diverse** utterances. On the one hand, redundant sentences are undesirable in the final output, on the other hand, they are useful for distinguishing more relevant and reliable descriptive utterances.

Table 6. Coverage of patterns.

	π_1^{es}	π_2^{es}	π_3^{es}	π_4^{es}	π_5^{es}
`Baseline ES-I`	78/37	17/10	00/00	13/10	05/03
`Mdef-WQA`	470/254	168/95	03/01	59/58	54/36

We considered an entirely different evaluation for each language for the following reasons: (a) the way the performance of definition QASs is measured differs between TREC and CLEF, and (b) CLEF gold standards for definition questions supply only

Table 7. Results overview. (TQ = Total number of questions in the question-set)

Corpus	Baseline EN-I			Mdef-WQA			
	TQ	AQ	NS	Accuracy	AQ	NS	Accuracy	AS (%)
(1)	133	81	7.35 ± 6.89	0.87 ± 0.2	**133**	18.98 ± 5.17	0.94 ± 0.07	16 ± 20
(2)	50	38	7.7 ± 7.0	0.74 ± 0.2	**50**	14.14 ± 5.3	0.78 ± 0.16	5 ± 9
(3)	86	67	5.47 ± 4.24	0.83 ± 0.19	78	13.91 ± 6.25	0.85 ± 0.14	5 ± 9
(4)	185	160	11.08 ± 13.28	0.84 ± 0.2	173	13.86 ± 7.24	0.89 ± 0.15	4 ± 11
(5)	152	102	5.43 ± 5.85	0.85 ± 0.22	136	13.13 ± 6.56	0.86 ± 0.16	8 ± 14

one nugget regarding abbreviations or position of persons, whereas TREC 2003 provides a set of relevant nuggets.

To start with the discussion of the obtained results, table 7 shows the coverage of `Baseline EN-I` and `Mdef-WQA`. AQ stands for the number of questions, for which its response contained at least one nugget (manually checked). `Mdef-WQA` discovered nuggets for all questions in (2), contrary to [1], who found nuggets for solely 42 questions by using external dictionaries and web snippets. In addition, `Mdef-WQA` discovered nuggets within **snippets** for the 133 questions in (1), in contrast to [10], who found a top five ranked snippet that conveys a definition solely for 116 questions within top 50 **downloaded full documents**.

Overall, `Mdef-WQA` covered 94% of the questions, whereas `Baseline EN-I` 74%. This difference is mainly due to the query rewriting step and the more flexible matching of δ. For all questions, in which `Mdef-WQA` and `Baseline EN-I` discovered at least one nugget, the accuracy and the average number of sentences (NS), containing also at least one nugget, was computed. `Mdef-WQA` doubles the number of sentences and achieves a slightly better accuracy. In table 7, AS corresponds to the percentage of sentences within NS, for which the relaxed matching shifted δ to another concept. Some shifts caused interesting descriptive phrases. A good example is: "*neuropathy*" was shifted to "*peripheral neuropathy*" and "*auditory neuropathy*", conversely, some shifts caused loosely related sentences: "*G7*" to "*Powershot G7*".

In order to compare our methods with a gold standard for English, we used the assessors' list provided through the TREC 2003 data. Following the approach of TREC, table 8 displays our current achievement. Given the higher recall 0.61 ± 0.33 obtained by `Mdef-WQA`, it can be concluded that the additional sentences that it selects contain more nuggets seen as vital on the assessor's list. A key point for the interpretation of the precision is the completeness of the assessor's list. It is known that systems in TREC are able find valid nuggets, which are judged as not relevant in the list (cf. [5] for details). This is even more likely for web-based system like `Mdef-WQA`, because they will discover many additional nuggets charged as relevant by a user, but will not hit the list. This kind of "it-is-not-on-my-list-evaluation" actually brings about a decrease, because they enlarge the response without increasing precision. In `Mdef-WQA`, this is a critical aspect, because it increases almost twofold the amount of selected descriptive sentences per question (see table 7), and hence, the length of the response.

Given the F(β) score achieved for each response by `Mdef-WQA` (see table 9) [14], it can be concluded: (a) it is "competitive" with the best systems in TREC 2003, which achieved between 0.5 and 0.56 for β=5, and (b) additional sentences provided novel nuggets.

Table 8. TREC 2003 results.

	Recall	Precision	Av. len.
`Baseline`	0.35 ± 0.34	0.30 ± 0.26	583
`Mdef-WQA`	0.61 ± 0.33	0.18 ± 0.13	1878

Table 9. TREC 2003 F(β) scores.

β	1	2	3	4	5
`Mdef-WQA`	0.26	0.37	0.45	0.50	0.53
`Baseline EN-I`	0.26	0.30	0.32	0.32	0.34

It is also worth to remark that `Baseline EN-I` obtained a slightly better F(β=5) for the following δs: "*Akbar the Great*", "*Albert Ghiorso*" and "*Niels Bohr*". This simply means that these responses were closer to the the assessors' expectations.

Table 10. Gold standards.

	`Baseline ES-I`	`Baseline ES-II`	`Mdef-WQA`
(4)	11	**33**	32
(5)	9	12	**22**

For Spanish, `Mdef-WQA` answered 32 and 22 out of the CLEF 2005 and 2006 questions respectively (see table 10). However, the runs submitted by the best two systems in CLEF 2005 answered 40 out of the 50 definition questions [13, 11]. Nevertheless, the third best system only answered 26 questions. Additionally, the best system in CLEF 2006 answered 35 out of the 42 definition questions, whereby `Mdef-WQA` found answers for 22 out of the 35 questions answered by this best system. Unfortunately, CLEF 2006 gold standard provides only one nugget for only these 35 questions.

Since the coverage of the gold standards focuses solely on abbreviations and positions of persons, and answers for seven CLEF 2006 questions are missed, we assigned three out of five different assessors to each data-set. Each assessor judged whether or not each output sentence yielded descriptive information. A sentence was considered as descriptive if and only if at least two out of the three assessors agreed (results in table 11). In both data-sets, `Mdef-WQA` outperformed both baselines, in particular, it discovered descriptive phrases for 47 out of the 50 CLEF 2005 questions. Additionally, `Mdef-WQA` returned more descriptive utterances (NS) with a lower level of redundancy. However, the accuracy of the output sentences decreased compared to our English results. We interpret this as a consequence of the lower amount of web redundancy for Spanish, which effects the quality of identifying the most relevant and reliable phrases. Finally, table 10 shows that the performance of `Mdef-WQA` can be improved by aligning patterns in [11] without necessarily considering them in the rewriting process.

All in all, the substantial difference in the performance between `Baseline EN/ES-I` and `Mdef-WQA` stresses the improvement caused by the query rewriting, and proves that extracting answers to definition questions straightforwardly from web snippets is promising.

Table 11. Results overview. (TQ = Total number of questions in the question-set)

Corpus		Baseline ES-I			Baseline ES-II		
	TQ	AQ	NS	Accuracy	AQ	NS	Accuracy
(4)	50	26	2.59 ± 2.45	0.85 ± 0.23	39	10.13 ± 10.66	0.67 ± 0.31
(5)	42	10	3.00 ± 3.13	0.61 ± 0.31	15	3.4 ± 3.31	0.65 ± 0.26

Corpus		Mdef-WQA		
	TQ	AQ	NS	Accuracy
(4)	50	**47**	8.6 ± 4.85	0.63 ± 0.19
(5)	42	**30**	7.27 ± 6.76	0.67 ± 0.25

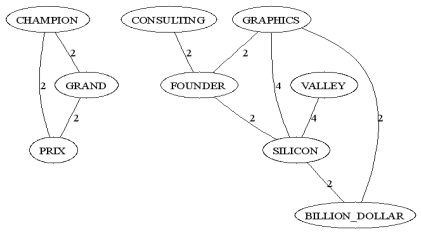

Fig. 1. $\hat{\Phi}_{ij} > 1$ for $\delta =$"*Jim Clark*".

Concerning the performance of the sense disambiguation process, `Mdef-WQA` was able to distinguish different potential senses for some δs, e.g., for "*atom*", the particle–sense and the format–sense. On the other hand, some senses were split into two separate senses, e.g., "*Akbar the Great*", where "*emperor*" and "*empire*" indicated different senses. This misinterpretation is due to the independent co-occurrence of "*emperor*" and "*empire*" with δ, and the fact that they are unlikely to share words. In order to improve this, some external sources of knowledge are necessary. This is not a trivial problem, because some δs can be extremely ambiguous like "*Jim Clark*", which refers to more than ten different real-world entities. `Mdef-WQA` recognised the pilot and the Netscape founder (Fig. 1). Independently of that, we found that entities and the correlation of highly closed terms in the semantic space provided by LSA can be important building blocks for a more sophisticated strategy for the disambiguation of δ.

4 Conclusions and future work

This work presents `Mdef-WQA`, a system that extracts answers for definition questions from web snippets. Our ongoing research focuses on adapting our system to deal with German. This adaptation brings about two challenges: (a) discriminate descriptive phrases in present tense from sentences in perfect tense with "*sein*", and (b) cope with the orthographical variations caused by umlauts and compounds.

`Mdef-WQA` pioneers attempts by definitional QAS to disambiguate descriptive utterances. One finding is that web snippets do not provide the necessary information for a complete disambiguation. To overcome this problem, external resources such as full documents, WordNet and/or additional queries might be explored as a source for fetching extra information from the web.

An additional challenge is recognising of relevant morpho-syntactical variations of descriptive sentences, which would help to decrease the redundancy of the output. Anyway, this redundancy can still be useful for discovering answers to definition questions in the context of the TREC/CLEF Question Answering tracks, projecting these redundant utterances to the corresponding corpus.

References

1. T.S.C.H. Cui, M.Y. Kan and J. Xiao. *A comparative study on sentence retrieval for definitional question answering*, SIGIR Workshop on Information Retrieval for Question Answering (IR4QA), July 29, 2004, Sheffield, UK.
2. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R.Harshman. *Indexing By Latent Semantic Analysis*, Journal of the American Society For Information Science, 41, 391–407, 1990.
3. C. Denicia-Carral, M. Montes-y-Gómez, L. Villaseñor-Pineda and R. García Hernández. *A Text Mining Approach for Definition Question Answering*, Lecture Notes in Computer Science, Volume 4139, pp. 76–86, 2006.
4. J. Goldstein, V. Mittal, J. Carbonell and M. Kantrowitz. *Multi-document summarization by sentence extraction*, NAACL-ANLP 2000 Workshop on Automatic summarization, pp. 40–48, 2000.
5. W. Hildebrandt, B. Katz and J. Lin. *Answering Definition Questions Using Multiple Knowledge Sources*, HLT-NAACL 2004, pp. 49–56, 2004.
6. H. Joho and M. Sanderson. *Large Scale Testing of a Descriptive Phrase Finder*, 1st Human Language Technology Conference, San Diego, CA, pp. 219–221, 2001.
7. H. Joho and M. Sanderson. *Retrieving Descriptive Phrases from Large Amounts of Free Text*, 9th ACM conference on Information and Knowledge Management, McLean, VA, pp. 180–186, 2000.
8. W. Kintsch. *Predication*,Cognitive Science 25, pp. 173–202, 1998.
9. B. Magnini, D. Giampiccolo, P. Forner, C. Ayache, P. Osenova, A. Peas, V. Jijkoun, B. Sacaleanu, P. Rocha and R. Sutcliffe. *Overview of the CLEF 2006 Multilingual Question Answering Track*, Working Notes for the CLEF 2006 Workshop, 20-22 September, Alicante, Spain, 2006.
10. S. Miliaraki and I. Androutsopoulos. *Learning to Identify Single-Snippet Answers to Definition Questions*, COLING 2004, pp. 1360–1366, 2004.
11. M. Montes-y-Gómez, L. Villaseñor-Pineda, M. Pérez-Coutiño, J. M. Gómez-Soriano, E. Sanchis-Arnal and P. Rosso. *INAOE-UPV Joint Participation in CLEF 2005: Experiments in Monolingual Question Answering*, Working Notes for the CLEF 2005 Workshop, 21-23 September, Vienna, Austria, 2005.
12. M. M. Soubbotin. *Patterns of Potential Answer Expressions as Clues to the Right Answers*, Proceedings of the TREC-10 Conference, NIST (2001), Gaithersburg, Maryland, 2001.
13. A. Vallin, D. Giampiccolo, L. Aunimo, C. Ayache, P. Osenova, A. Peñas, M. Rijke, B. Sacaleanu, D. Santos and R. Sutcliffe. *Overview of the CLEF 2005 Multilingual Question Answering Track*, Working Notes for the CLEF 2005 Workshop, 21-23 September, Vienna, Austria, 2005.
14. E. M. Voorhees. *Evaluating Answers to Definition Questions*, HLT-NAACL 2003, pp. 109–111, 2033.
15. P. Wiemer-Hastings and I. Zipitria. *Rules for Syntax, Vectors for Semantics*, Proceedings of the 23rd Annual Conference of the Cognitive Science Society, 2001.