13

Ranking and Extraction of Relevant Single Words in Text

João Ventura and Joaquim Ferreira da Silva
DI/FCT Universidade Nova de Lisboa
Portugal

1. Introduction

The extraction of keywords is currently a very important technique used in several applications, for instance, the characterization of document topics. In this case, by extracting the right keywords on a query, one could easily know what documents should be read and what documents should be put aside. However, while the automatic extraction of multiword has been an active search field by the scientific community, the automatic extraction of single words, or unigrams, has been basically ignored due to its intrinsic difficulty. Meanwhile, it is easy to demonstrate that in a process of keyword extraction, leaving unigrams out impoverishes, in a certain extent, the quality of the final result. Take the following example:

> *The budgets have deteriorate due to the action of automatic stabilisers and also because the discretionary fiscal expansionary measures of some Member-States who had no room for manouvre. In general, and despite budgetary pressures, public investment has remained static or increased slightly, except in Germany, Greece and Portugal.*

According to the previous example, one can easily identify several relevant terms. But, if in one hand, multiword terms such as "automatic stabilisers", "discretionary fiscal expansionary measures ", "budgetary pressures" and "public investment" would be easily captured by the modern multiword extractors, uniword terms like "budgets", "Member-States", "Germany", "Greece" and "Portugal" would not. However, a simple count demonstrates that in this example there are almost as many multiword as uniword terms. In fact, the relevant unigrams of a document are usually part of the important topics in it, as it may also be the relevant multiwords, and in the previous example, terms such as "Germany", "Greece" and "Portugal" should be considered extremely important because they are names of countries.

In this chapter we will look into the problematic of unigram extraction, reviewing some of the current state-of-the-art techniques and comparing its results with two metrics proposed by us. We'll also review a new technique proposed by us based on the syllable analysis that is able of improving the results in an unorthodox way. Finally, we shall present the "Islands method", a technique also proposed by us that allows one to decide about the *boolean* relevancy of a certain word.

2. Current state-of-the-art methods for unigram extraction

The current state-of-the-art approaches can be subdivided into several groups. On one side we have the linguistic approaches. These approaches are able to extract information from documents using linguistic information (morphological, syntactic or semantic) about a given text. Usually, this kind of information is obtained from the use of grammars like in (Afrin, 2001), or from annotated texts. Although there are reliable automatic techniques for text annotation or grammar building, those kinds of approaches are usually language dependent, making the generalization of such methods a very difficult aim. Other linguistic approaches, like the one used in (Heid, 2000), extracts relevant terms using regular expressions. However, in that work the author, by using 20 prefixes and 20 suffixes carefully extracted from German language shows how dependent from the language his method is.

In the same line we have the approaches based on knowledge. Those approaches are usually associated with ontologies where the main idea is to get a representative model of the specific reality of the analyzed documents. A simple example for extraction of relevant information using knowledge based approaches can be associated with the knowledge of the structure of documents to, for instance, extract keywords from the titles and abstracts of scientific documents. More complex examples, like (Gao & Zhao, 2005), are able to identify frauds on emails. However, these kinds of approaches are also quite limiting, mainly because the creation of ontologies isn't straightforward, and ontologies are something very specific to a certain subject and can't be easily generalized. For instance, in the case of keyword extraction from titles of scientific texts, one has to know exactly the structure of those documents in order to identify where the titles and abstracts are. On the other hand it's almost impossible to use those kinds of methods on documents without apparent structure.

Other authors have also tried to use Neural Networks to do unigram extraction. A Neural Network is a programming model that resembles, in a certain way, the biological neural model. Applied to information extraction, the most common application is based on a user's query answering. Made simple, a user queries a set of documents and the neural net verifies if the user query is relevant in a certain document or not. If it is, that document is retrieved and presented to the user. In (Das, 2002) a technique based on Neural Networks is presented. The basic idea is that each of the nodes (or neurons) has a user's query word associated with it. For each word on an input scientific paper, the nodes which have query words that exist on the input paper are raised to a higher level of *energy*. This process continues until the neural network stabilizes. From this, one can see which nodes have higher energy levels for that document and thus, more relevant to the query. However, also this kind of approach has problems. Neural Networks are usually slow while building because of backpropagation calculations. In this way, a neural net handling 15.000 words, the average size of a single scientific paper, or 700 distinct words would be too slow, considering you would have to create a neural network each time a user makes a query, multiplying it for the amount of documents where the user would want to search in.

Finally, following the same line as the previous ones, we also have the hybrid approaches that aim to bring the best of all the other into a single one. In (Feldman et al., 2006) the authors are using grammars in conjunction with statistical methods in order to extract information from web pages and convert them to semantic web pages. In that paper the rules of the grammar used were manually created and the probabilities used were extracted

from an annotated corpus. Also in this case there is overdependence again on something: the annotated corpus and the manual creation of the grammar.
At last, following a different line than the previous methods, we have the statistical based approaches. The main advantages in those kinds of approaches are the faster implementation and usage of the methods and the independence in relation to the language used on the texts, in relation to the structure used and to the context of the documents tested. In the next three subsections we will review three of the most known statistical approaches for information retrieval: Luhn's frequency criterion, Tf-Idf method and Zhou's & Slater method.

2.1 Luhn's frequency criterion

Luhn, in one of the first published papers concerning relevant unigram extraction techniques (Luhn, 1958), suggests a method for the classification of unigrams based on the frequency of ocurrence of terms. According to the author,

> *"… the justification for measure the relevance of a word by the frequency of ocurrence is based on the fact that a writer usually repeats some words when arguing and when elaborates certain aspects of a subject...."*

Luhn also suggested that the words with a very high frequency of ocurrence are usually considered common words and unfrequent words could be considered rare, both cases being unrelevant words. Although this approach seems quite intuitive, is not necessarily true. During our research with corpora of different languages, among the 100 more frequent words, in average, about 35% could be considered relevant. Table 1 lists some of those words:

Word	Rank	Frequency
Comission	28	1909
Member-States	38	1378
Countries	41	1219
European	55	874
Union	92	515
Europe	99	463

Table 1. Some words among the 100 more frequent ones in an English corpus

Considering the fact that in average the corpora used in our work has about 500.000 words, from which about 24.000 are distinct, one can easily understand that with this criterion possibly some or all of the words listed in table 1 would be thrown away. Luhn's criterion becomes, in this case, quite restrictive. And if we consider the fact that the words in table 1 came from European Union texts, one can see the kind of the information that would be rejected. Words like "European" and "Union" are preety descritive of the texts.
Other problem with this approach has to do with the thresholds. How can one find the threshold between very frequent words and relevant words? Or between the relevant words and rare words? Finally, Luhn considers that the relevant words are those not very frequent nor very rare. Again, this may be a problem because not all of the words between those thresholds are important. Luhn solves partially this problem using a list of common words that should be reject on the final list. But Luhn idealized its method for texts with an average

of 700 distinct words (scientific papers) and it would be impracticable to mantain a list of common words handling texts with 24.000 distinct words.

2.2 Tf-Idf

Tf-Idf, Term Frequency – Inverse Document Frequency (Salton & Buckley, 1987), is a metric for calculating the relevance of terms in documents, very used in Information Retrieval and Text-Mining. Essentially, this technique measures how important a certain word is on a document regarding other documents in the same collection. Basically, a word gets more important in a certain document the more it occurs in that document. But if that word occurs in other documents, its importance decreases. Words that are very frequent on a single document tend to be more valued than common words that occur on more documents, like articles or prepositions.

The formal procedure for the implementation of Tf-Idf changes slightly from application to application, but the most common approach was the one used in this work. Generally, the calculation of Tf-Idf is made in separate, calculating the Tf and Idf components separately, and finally multiplying both components to get the final Tf-Idf value.

Tf component (term frequency) simply measures the number of times a word occurs on a certain document. That count is then normalized to prevent word on very long documents to get higher Tf values. Equation 1 measures the probability that a term i occurs in a document j.

$$Tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \tag{1}$$

where $n_{i,j}$ is the number of times the term i occurs in a document j and then it is divided by the total of words in document j.

Idf component measures the general relevance of a given term. Equation 2 consists in the count of the number of documents that a term t_i occurs.

$$Idf_i = \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}, \tag{2}$$

where $|D|$ represents the total number of documents in the collection and $\{d_j : t_i \in d_j\}$ the number of documents where the term t_i occurs.

Tf-idf (equation 3) is then the multiplication of the two previous equations.

$$TfIdf_{i,j} = Tf_{i,j} * Idf_i. \tag{3}$$

However, we must consider that the main goal of this method is to analyze the relevance of a word in a document regarding other documents, instead of analyzing the relevance of a word in corpora. To do that, we had to change slightly the method. Basically, and because the corpora used for research were made from single documents, we've adapted the method to give a word the maximum Tf-Idf found in all methods. In this way, we can use Tf-Idf to evaluate a word's relevance on corpora.

Unfortunately, also Tf-Idf has problems. Similarly to Luhn's frequency criterion, Tf-Idf harms the very frequent relevant words because they tend to exist in almost all documents, and so, the Idf component lowers the final Tf-Idf value. On the other side, the Idf component also damages certain words by not taking into account the probabilities of

occurrence of a word in other documents. For instance, if you have three documents, and a certain word occurs 100 times in one document, and just once in the other documents, the Idf component gets equal to zero when it's pretty clear that that word is, probably, very relevant in the document where it occurs 100 times. If that same word occurs 1 or 50 times in the other two documents it's almost irrelevant to Tf-Idf, but however, occurring 1 or 50 times in those two other documents means different things about that same word.
At last, the Idf component also has the problem of benefiting rare words because if, for instance, in a document exists a unique orthographical error, it gets the maximum Idf value available.

2.3 Zhou & Slater method

Zhou & Slater method is a very recent metric proposed in (Zhou & Slater, 2003) for calculating the relevance of unigrams. It is assumed, in some way similarly with Tf-Idf and Luhn's criterion, that the relevant words can be found in certain areas of the texts either by being part of the local topics, either by being related to the local contexts, therefore forming clusters in those areas. On the other hand, common and less relevant words should occur randomly in all the text, therefore not forming significant clusters.
This technique, being an improvement and extension over the technique proposed in (Ortuño et al., 2002) measures the relevance of a word accordingly to the position of occurrence of each word in texts.
Starting with a list L_w={-1, t_1, t_2, …, t_m, n}, where t_i represents the position of the i-th occurrence of the word w in the text and n represents the total number of words in the same text, we obtain $\hat{u}$ that is basically the average separation between successive occurrences of word w.

$$\hat{u} = \frac{n+1}{m+1}. \tag{4}$$

Next step consists in the calculation of the average separation of each occurrence of the word w, using equation 5.

$$d(t_i.) = \frac{t_{i+1} - t_{i-1}}{2}, \qquad i = 1, \ldots, m. \tag{5}$$

On equation 4 we have the average distance between all successive occurrences of word w. With equation 5 we get the local information for each point t_i, meaning that we get the average separation between each occurrence of the word w in the text.
The next step consists in the identification of the points on L_w which form part of clusters. Basically a point forms part of a cluster if its average distance $d(t_i)$ (average distance between the previous and next occurrence of the same word) is less than the average distance between occurrences ($\hat{u}$). In this way, we get $\delta(t_i)$ which, according to equation 6, identifies which points t_i belong to clusters.

$$\delta(t_i) = \begin{cases} 1, & if\ d(t_i) < \hat{u} \\ 0, & otherwise \end{cases}. \tag{6}$$

In a parallel way, using equation 7 we get $v(t_i)$ that represents the local excess of words relating position t_i. It basically consists in the normalized distance to the average value of distance.

$$v(t_i) = \frac{\hat{u} - d(t_i)}{\hat{u}}. \quad (7)$$

By equation 7, the less the value of $d(t_i)$ (or the closer the t_i points are), the bigger the value of $v(t_i)$ because, as stated before, the purpose of this technique is to value the formation of clusters.

$$\Gamma(w) = \frac{1}{m} \sum_{i=1}^{m} \delta(t_i) * v(t_i). \quad (8)$$

Therefore, starting from the list L_w={-1, t_1, t_2, …, t_m, n}, we get the score of the word w using equation 8. Being in $\delta(t_i)$ the information about whether t_i belongs or not to a cluster, and in $v(t_i)$ the normalized distance to the average distance, $\Gamma(w)$ gets the value of $v(t_i)$ when t_i belongs to a cluster and the value of zero otherwise.
Although this is a very efficient and ingenious method, it has also the same problems as the previous ones regarding the very frequent relevant words. In a general way, all the methods that assume that relevant words occur only in certain areas of the texts suffer from that problem. Although there is a certain veracity in it, it damages the very frequent relevant words because they tend to occur allover the text and not only on local contexts. Also, by dealing exclusively with significant clusters, the relevant words with low frequency of occurrence are also very damaged by this method.

3. An alternative contribution

In this section we will present a set of innovative alternatives to the previous presented methods. We will present two new metrics recently proposed by us (Ventura & Silva, 2007) for the calculation of the relevance of unigrams, the measure Score and SPQ. We will also present a new research field based on the syllable analysis of the words and finally we will present a new unigram extractor that we've called "Islands Method".

3.1 A word about relevance

Starting with a corpus composed of several documents, one of the objectives of this work is to try to understand which words are relevant and which words are not. However, using purely statistical methods, this kind of classification isn't always straightforward or even exact because, although the notion of relevance is a concept easy to understand, normally there's no consensus about the frontier that separates relevance from non-relevance. For instance, words like "Republic" or "London" have significative relevance and words like "or" and "since" have no relevance at all, but what about words like "read", "terminate" and "next"? These kind of words are problematic because usually there's no consensus about their semantic value. So, there is a fuzzy frontier about the relevance of words. In this way, regarding the context of this work, we've decided to adopt a conservative approach and classify as relevant only those words with inquestionable semantic value.

3.2 The Score measure

One of the first steps for the extraction of relevant unigrams consists in obtaining a list ranked by the potential relevance of each of the words in a corpus. This list measures therefore the relative relevance of each word regarding other words occurring in a corpus, so, a word ranked higher in the list is considered more relevant than a word occurring in the bottom of the list. To do this we've developed a new metric where the main idea is that the relevant words usually have a special preference to relate with a small group of other words. In this way, it is possible to use a metric that measures the importance of a word in a corpus based on the study of the relation that that word has with the words that follow it. We have denominated that measure the successor's score of a word w, that is $Sc_{suc}(w)$.

$$Sc_{suc}(w) = \sqrt{\frac{1}{||\gamma|| - 1} \sum_{y_i \in \gamma} \left(\frac{p(w, y_i) - p(w, .)}{p(w, .)} \right)^2}. \tag{9}$$

In equation 9, γ is the set of distinct words in the corpus and $||\gamma||$ stands for the size of that set; $p(w\ y_i)$ represents the probability of y_i to be a successor of word w; $p(w,.)$ gives the average probability of the successors of w, which is given by:

$$p(w, .) = \frac{1}{||\gamma||} \sum_{y_i \in \gamma} p(w, y_i) \qquad p(w, y_i) = \frac{f(w, y_i)}{N}, \tag{10}$$

where N stands for the number of words occurred in the corpus and $f(w, yi)$ is the frequency of bigram (w, yi) in the same corpus. Resuming the mathematical formalism, $Sc_{suc}(w)$ in equation 9 is given by a standard deviation *normalized* by the average probability of the successors of w. It measures therefore the variation of the current word's preference to appear before the rest of the words in the corpus. The higher values will appear for the words that have more diversified frequencies with the words that follow it, and the lowest values will appear in the words that have less variations of frequency with words that follow it. Similarly, we measure the preference that a word has to the words that precede it using the following metric that we've denominated predecessor's score, that is $Sc_{pre}(w)$.

$$Sc_{pre}(w) = \sqrt{\frac{1}{||\gamma|| - 1} \sum_{y_i \in \gamma} \left(\frac{p(y_i, w) - p(., w)}{p(., w)} \right)^2}, \tag{11}$$

where the meanings of $p(y_i,w)$ and $p(.\ ,w)$ are obvious.

So, using both equations 9 and 11 through the arithmetic average, we will obtain the metric that allows us to classify the relevance of a word based on its predecessors and successors. This metric is simply denominated $Sc(w)$.

$$Sc(w) = \frac{Sc_{pre}(w) + Sc_{suc}(w)}{2}. \tag{12}$$

It can be seen by the previous expressions that Score measure gives better values to a word that as the tendency to attach to a restricted set of successor and predecessor words. However, it can be easily noted that this metric benefits extremely the word with the frequency of 1, because when a unigram occurs only once in a corpus, the relation with its successor and predecessor is unique, or in other words, complete. In this way, Score interprets that relation as a strong correlation, and so care must be taken to pre-process the corpus in order to remove the unigrams with frequency 1. This situation doesn't mean that frequency affects directly results; the correlation in the cases of frequency 1 is effectively high

and that occurs because we're using a standard deviation. In any statistical approaches, higher frequencies represent better reliability on the results quality. For low frequencies it can be assumed that the results, whatever they are, can't be considered statistically conclusive. Table 2 shows some examples of *Sc(.)* values and ranking positions for the words of an English corpus made from documents of the European Union. It has about half million words and there are 18,172 distinct ones. We've studied the words that occur at least 3 times in the corpus. As one can see, the more common words like "the", "and" and "of" are positioned lower in the ranking while words with semantic value are positioned upper in the list.

Word	*Sc (.)*	Rank
pharmacopoeia	135.17	48
oryctolagus	134.80	64
embryonic	132.67	76
of	24.15	6627
the	19.34	6677
and	10.82	6696

Table 2. Some examples of *Sc(.)* values and ranking positions for words in an English corpus

3.3 SPQ measure

By observing some characteristics of the unigrams, it was also verified that the words considered relevant usually have some interesting characteristics about the number of predecessors and successors. For instance, with a Portuguese corpus of half million words (also from European Union documents), it could be noted that the relevant word "comissão" (commission) occurred 1.909 times in the corpus, with 41 distinct predecessors and 530 distinct successors. Also, the relevant word "Europa" (Europe) occurred 466 times in the corpus, with 29 distinct predecessors and 171 distinct successors. In both cases, most of the predecessors are articles or prepositions such as "a", "na" e "da" (the, on and of). In fact, function words (articles, prepositions, etc.) show no special preference to a small set of words: one may say that they populate the entire corpus.

The morphosyntactic sequence <article> <name> <verb> is very frequent in the case of Latin languages such as Portuguese, Spanish, Italian and French, among others. In these cases, given that there are more verbs than articles, it is natural that names have more successors than predecessors. Looking at table 3, we can find some examples of morphosyntactic sequences, and note that the list of articles is usually small while the list of verbs is more extensive.

Morphosyntactic sequences
a comissão lançou
a comissão considera
a comissão europeia
pela comissão tratada

Table 3. Example of morphosyntactic sequences in Portuguese

Following this reasoning, we have proposed another statistic metric that measures the importance of a word based on the quotient between the number of its distinct successors and the number of its distinct predecessors. We have called it *SPQ* (Successor-Predecessor Quotient).

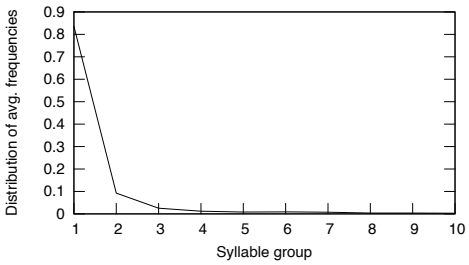
$$SPQ(w) = \frac{Nsuc(w)}{Nant(w)}, \tag{13}$$

where *Nsuc(w)* and *Nant(w)* represent the number of distinct successors and predecessors of word *w* in the corpus.

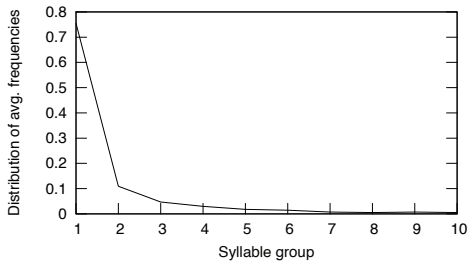
However, although both presented metrics (Sc and SPQ) measure the relevance of words, in a language-independent basis, when we tested SPQ, the results were better for the Portuguese and Spanish corpora than for the English one. However, assuming this, it may be preferably to use this metric if one is working only with Latin languages (see results in section 4).

3.4 Syllable Analysis

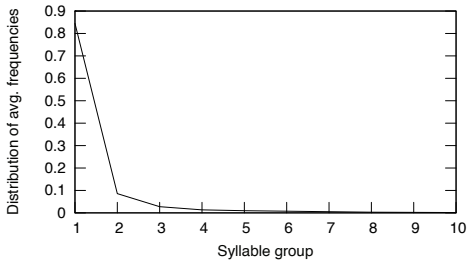
Considering again table 2 in section 3.2, one can find that from those 6 words, 3 are relevant and 3 are not. It is easy to conclude that the relevant words ("pharmacopoeia", "oryctolagus" and "embryonic") are, in fact, larger than the non-relevant ("of", "the" and "and"). We could build a metric in order to favour larger words as they appear to be more relevant, but, as we will see, it is preferable to consider the number of syllables instead of the length of the words. For instance, the probability of occurrence of the definite article "the" in oral or textual speeches is identical to its Portuguese counterpart article "o". However, there is a 3-to-1 relation about the number of characters, while the number of syllables is identical in both languages (one syllable). Thus, a metric based on the length of words would value the word "the" 3 times more relevant than the word "o", which wouldn't be correct. Using a metric based on the number of syllables, that distortion would not occur.



(a) Portuguese Corpus



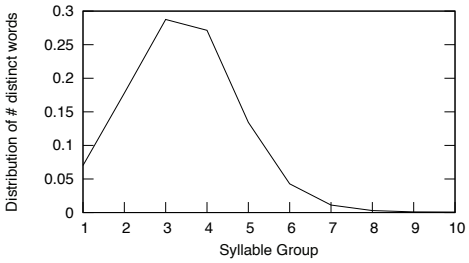
(b) English Corpus



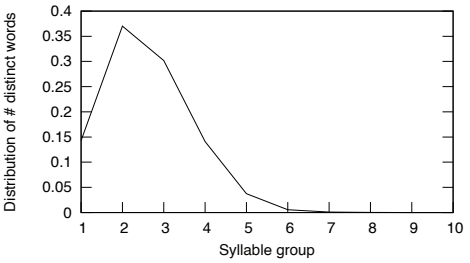
(c) Spanish Corpus

Figure 1. Normalized distribution of the average frequency of words occurrence for each syllable group, for all the three researched corpora

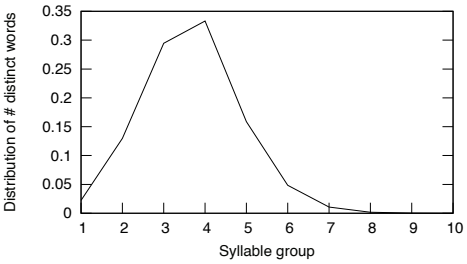
Figure 1 shows the distribution of the average frequency of words occurrence for each syllable group for all the corpora researched: Portuguese, Spanish and English; the values are normalized such that its sum is 1. Each one of the graphics in figure 1 represents, basically, the average frequency of occurrence of the words belonging to each syllable group, i.e., having that exact number of syllables. Looking at those graphics it is possible to see that the words with one syllable occur more frequently than the words with two syllables, followed by the words with two syllables, etc. So, the average frequency of occurrence of the words in each syllable group decreases with the increase of the number of syllables. This phenomenon is certainly related to the economy of speech. It is necessary that the words that occur more often are the ones easier to pronounce, otherwise the discourses would be too long. The words having 1 syllable are usually articles and other function words like "and", "the", "of" and "or" (in Portuguese "e", "o", "de" and "ou"); because they occur more frequently in texts, they must be easier and faster to pronounce.



(a) Portuguese Corpus



(b) English Corpus



(c) Spanish Corpus

Figure 2. Normalized distribution of the number of distinct words for each syllable group, for all the three researched corpora

Figure 2 shows the distribution of the number of distinct words for each syllable group, for the English, Portuguese and Spanish corpora; the values are normalized such that its sum is 1. The interpretation of this curve is beyond the domain of this work, but without a secure certainty, we believe that these distributions are probably connected to the number of distinct words that may be formed preferably with the least number of syllables, considering the legal sequences that may be formed in each language. In fact, the number of words that may exist with 2 or more syllables is certainly greater than the number of words with 1 syllable. In the Portuguese case, for instance, the maximum peak occurs in the 3 syllables group, while in the Spanish case the peak occurs in the 4 syllable group and the English peak in the 2 syllable group. This is probably because the Portuguese language is

usually more restrictive than the English language concerning the possible number of character combinations for each syllable, needing to occupy the 3 syllables group. The same can be said regarding the Spanish corpus and the 4 syllable group. Another possible explanation for this phenomenon can be related to table 4, which shows the average number of letters of the words in each syllable group. In this way, we can see that, in average, the English words with 1 syllable have 4.7 letters while the Portuguese and Spanish words with 1 syllable have, respectively, 3.7 and 3.9 letters.

Corpus	1-S	2-S	3-S	4-S	5-S	6-S	7-S	8-S	9-S	10-S
Portuguese	3.7	5.5	7.6	9.7	11.8	14.0	16.2	18.4	21.1	27.0
English	4.7	6.8	8.9	10.8	12.9	15.5	18.8	23.3	22.0	24.0
Spanish	3.9	5.6	7.5	9.5	11.5	13.6	15.7	18.3	20.6	22.0

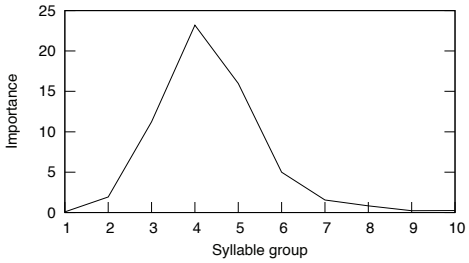
Table 4. Average number of letters for each syllable group for all the researched corpora

Also, according with table 4 the English language has in average more letters on the 8 first syllable groups than the other two languages. If it has more letters per syllable, it is natural that more combinations can be made with less syllables and maybe that is why the English languages reaches its peak before the other two languages. The Spanish and Portuguese languages have the same kind of graphic on the first two syllable groups and a slight inversion on the 3 and 4 syllable group which, besides language restrictions, can also be explained by the data in table 4.

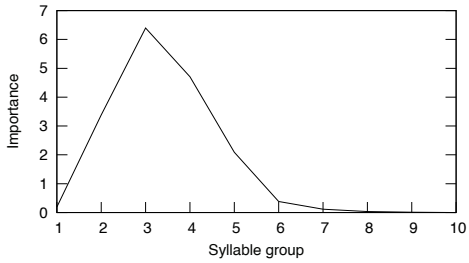
Thus, figure 2 shows us that in the case of the English language (the other languages can be analysed in a similar way) there is more diversity of words with 2 syllables. In the 1-syllable group we can find, above all, function words like articles and prepositions where there is no semantic value. On the other side, very rare words, with many syllables, have semantic contents which are too specific to be considered relevant and broad simultaneously. In the case of the Portuguese and Spanish languages they have their peak respectively in the 3-syllable group and 4-syllable group. Still, both Portuguese and Spanish graphics are quite similar which reflects the fact that both languages are descendent from a common language.

Figure 3 shows us three graphics that represents the importance of each syllable group for each language. For each syllable group, importance is determined by the corresponding values used in the graphics of figure 2 (the Normalized distribution of the number of distinct words) divided by the corresponding value used in the graphics of figure 1 (Normalized distribution of the average frequency of words occurrence). If the distributions on figure 3 were used to classify words on texts, the 4 syllable group for the Portuguese and Spanish case and the 3 syllable group for the English case would be the most important group, following by the other groups accordingly to the distributions.

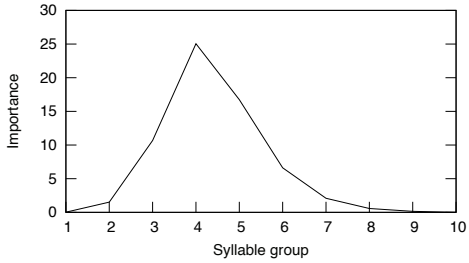
Although this method appears at first sight to be language dependent as it deals with very specific linguistic information, in fact it is not; that would be very disadvantageous because we want the methods to be as independent from any factors as possible . However we must mention that all the necessary information to obtain the previous distributions can be obtained directly from the research corpora. This way, if a corpus is sufficiently representative of a language, syllable distributions can be obtained, independently of the language.



(a) Portuguese Corpus



(b) English Corpus



(c) Spanish Corpus

Figure 3. Importance of each syllable group, for all the three researched corpora

3.5 The Islands method – A unigram extractor

Although all the previous methods presented here (including the ones stated in section 2 – state-of-the-art techniques) are capable of identifying to a certain extent the relevant words in texts or corpora, they are, however, incapable of making decisions about the true relevance of words. The problem is that all the previous methods can only create relevance rankings, from which we can only identify, for instance, that a certain word on the top of the list must be more relevant than a word on the bottom. However, in certain situations it may be necessary to know if a given word is truly relevant (like Boolean true or false) instead of knowing that this word is more relevant than X, Y and Z. This kind of certainty is absolutely necessary for applications like Documents-ID where we desire for a set of words that truly describes a document or set of documents.

On a first analysis one could consider that all the words on top of the ranking are relevant, and that the words on the bottom are not. But this causes two kinds of problems. The first is to define the frontier that separates the relevant from the non-relevant words. Where should this frontier be? If it is too high in the list, we'd probably miss relevant words to non-relevance. If it was too low it would be the opposite. The other problem is that although the words in the ranking can be generically compared among each other, i.e, we can say that a certain word X in the top of the rank is more relevant than a certain word Y in the bottom, we can't say that Y is not relevant at all even if it is at the very bottom of the list. This is because while X has a high score of relevance and so must be very relevant in all the text, Y may be very relevant only in a local context, getting a smaller score therefore being in the final of the *general* relevance list. This can even mean that Y is truly relevant in a local context while X is not relevant in the global context!

As far as we know, there is no such method to extract relevant words on this kind of basis. We present a method that we have designated "Islands method" which allows us to extract relevant words from a text, based on a relevance ranking previously generated.
Following the same line of idea as the Score method, the main assumption of the Islands methods is that a word, to be considered relevant, must be more important than the words in its neighbourhood. This means that each word is tested in its local context, whether this context is a paragraph or even the entirety of a text. Then, recurring to the relevance rankings given by the previous methods we are able to compare the importance of all the words in a text.
In our approach we start by considering the weight that each neighbours of a word has in terms of frequency. The idea is that the more a certain word co-occurs with another, the more important that connection is. We then proceed to the calculation of a weighted average, based on the frequency of co-occurrence. Then, if a word has its score greater than 90% of the weighted average of its neighbours, we assume it as relevant. Equations 14 and 15 measure the weighted averages of, respectively, the predecessors and successors of a word.

$$Avg_{pre}(w) = \sum_{y_i \in \{predecs\ of\ w\}} p(y_i, w) * r(y_i) \quad , \tag{14}$$

$$Avg_{suc}(w) = \sum_{y_i \in \{succecs\ of\ w\}} p(w, y_i) * r(y_i) \quad , \tag{15}$$

where $p(y_i,w)$ means the probability of occurrence of the bigram (y_i,w) and $r(y_i)$ is the relevance value given by the generic $r(.)$ metric. The same must be considered for equation 15. Thus, accordingly to the Islands criterion, a word w is considered relevant *if and only if*:

$$r(w) \geq 0{,}9.max(Avg_{pre}(w), Avg_{suc}(w)) \; . \tag{16}$$

As it shall be presented in the next section, the results for this method are very encouraging. Words which are somehow "isolated" in terms of score in the relevance rankings in relation with its neighbours are easily considered relevant. Words that are part of relevant *n-grams* (bigrams, trigrams, and so on) aren't easily excluded because of the 90% factor on the criterion.

4. Results

In this section we present the results concerning all the previous mentioned methods and techniques including the ones stated in section 2, state-of-the-art. We will briefly describe the used corpora, as well as the criterion used to evaluate the quality of the rankings generated by the methods. Then we shall present the results concerning the unigram extractor (Islands method) and at last we shall analyse the application of the syllable method over the techniques and metrics discussed.

4.1 The test corpora

The corpora used in this work, as already mentioned, are composed of several documents extracted from the Portal for the Access to the European Union law (http://eur-lex.europa.eu/). In this site we can find an enormous repository of documents and communications of public interest in the domain of the European Union.

We have extracted documents in three different languages and created three different corpora. The Portuguese corpus is made of 43 documents, and has about half million words from which about 24.000 are distinct. The English corpus is made of 40 documents, having also about half million words, from which about 18.000 are distinct. The Spanish corpus is made of 41 documents; it has about 550.000 words, from which 22.000 are distinct.

4.2 Test sets

The test sets are subsets of the tested corpora from which certain words are classified as relevant and other as irrelevant in order to test the quality of the methods listed in this work. Table 5 lists, for convenience, the several test sets and their description.

Test Name	Description
A	100 more frequent words.
B	200 random words from the 1.000 more frequent ones.
C	300 random words from the 3.000 more frequent ones.
D	200 random words with frequency of ocurrence greater than 1.
E	Includes all the previous ones.

Table 5. List of test sets and their description

Although the "D" test set seems sufficient to evaluate the efficiency of the metrics because it uses a set of words independent from the frequency, the other tests serve to add information about the behaviour of all the metrics in specific areas of frequency. Thus, with the test set "A" we pretend to evaluate the efficiency on the very frequent words, where on the contrary to the common sense, we can find several relevant words pretty illustrative of the corpora general topics. With test sets "B" and "C" we pretend to evaluate the metrics on the intermediate areas of frequency. With test set "D" we pretend to have a broader view of the methods ignoring words of frequency 1 that are usually orthographical errors. Finally "E" test aims to evaluate the metrics in a even broader view with a higher percentage of frequent words.

4.3 Evaluation criterion for relevance rankings

As mentioned previously, when considering relevance rankings of words there is a fuzzy area of relevance where, in a certain way, the relevance of certain words may be considered dubious. We've chosen to follow a conservative approach, considering relevant only those words that are unquestionable relevant. That said, after obtaining the relevance ranks the task was to evaluate their quality. Although this seems something very simple, the fact is that we didn't find any published approach to do this. For example, on the papers we had to research, although those authors have dealt with relevance rankings, it doesn't seem they have quantified their quality results, only showing the rank position of some words. For this work, we had to create a new method to quantify the quality of a relevance ranking. It is made in the following way: first, an evaluation of a certain method has to be made. After that we have the relevance ranking ordered by score. Then the following criterion is applied: if all the words manually considered relevant are in the top of the rank, there's 100% of efficiency. On the other hand, if none is on the top of the list, we get 0% of efficiency. If, for example, we have 30 relevant words, but only 25 of them are in the 30 first positions of the ranking, we get an efficiency of $25/30 \approx 83.3\%$. In the case when the number of relevant words is greater than the number of irrelevant words, we invert the case: instead of measure

the quality by the number of relevant words in the top, we find the number of irrelevant words scored low in the ranking. If all the irrelevant words are at the bottom, it means that the relevant words are in the top. For instance, with a test set of 100 words, if 90 of them are relevant and 10 irrelevant, if we only count the number of relevant words in the first 90 positions, we get efficiencies from 100% to a minimum of 89% (= 80/90). But if we invert the analysis, if we count the number of irrelevant words in the 10 bottom positions we can get efficiencies from 100% (when all the irrelevant words are in those 10 bottom positions) to 0% (when all the irrelevant words **are not** in those 10 bottom positions).

4.4 Precision and Recall for the Islands method

Precision and Recall are two statistical measures which allows to evaluate the quality of results in domains such as Information Recovery or Statistical Classification. Both these metrics deal with binary data. In this work they serve to obtain quantitative information about the quality of the unigram extractor (the Islands method). Their expressions are given in equations 17 and 18.

$$Precision = \frac{\#(relevant\_words \cap considered\_relevant)}{\#considered\_relevant} \quad . \tag{17}$$

$$Recall = \frac{\#(relevant\_words \cap considered\_relevant)}{\#relevant\_words} \quad , \tag{18}$$

where *relevant_words* is the set of words classified manually as relevants, *considered_relevant* is the set of words considered relevant by the unigram extractor and #*(relevant_word* ∩ *considered_relevant)* is the number of words that are relevant and were considered relevant by the extractor. Briefly, **Precision** measures the proportion of how many words considered relevant by the extractor are, in fact, really relevant, while **Recall** measures the proportion of really relevant words that were considered relevant by the extractor. For instance, if you have a test set where 100 words are really relevant and the extractor has only considered relevant one single word, if you only take the Precision measure, you'd get 100% of precision. This only means all the words considered relevant by the extractor are truly relevant. But if you'd take the value of Recall, you'd get a recall of 1%, and this would mean that although the extractor is correct in the extraction, it is pretty inefficient because only one of the 100 relevant words were considered relevant by the extractor. So, as we can see, both measures are important and inseparable.

4.5 Results

The following tables (tables 6 to 8) represent the results of quality of the several test sets presented in section 4.2, when applied to the methods presented as state-of-the-art (*Tf-Idf* and *Zhou & Slater* methods) as well as to the method proposed by us (*Score* and *SPQ*). We also present results for the syllable method isolated, i.e., as if it was a metric on its own for the evaluation of relevance, only by testing the number of syllables of each word in the test sets, getting the results accordingly to the importance of each syllable group (see figure 3). Finally we also present the results of applying the syllable method in conjunction with the other methods. The application of the syllable method to another metric is something straightforward: for each word and each *standard* metric (*Score*, *SPQ*, *Tf-Idf* and *Zhou & Slater*), we multiply the obtained score with the importance of its syllable group according to

its language and the correspondent graphic on figure 3. If a word is stated in those graphics as more important because of its number of syllables, the result after multiplying benefits it. Otherwise it gets the correspondent result.

Method	Test "A"	Test "B"	Test "C"	Test "D"	Test "E"
Syllables isolated	78.6	74.0	53.8	63.1	68.6
Sc	60.7	61.0	58.3	38.5	58.1
Sc & Syllables	85.7	79.2	57.5	63.1	69.0
SPQ	71.4	63.6	65.2	38.5	63.7
SPQ & Syllables	89.3	77.9	63.6	63.1	71.3
Tf-Idf	46.4	54.5	63.6	47.7	56.8
Tf-Idf & Syllables	78.6	76.6	62.1	60.0	68.0
Zhou	25.0	58.4	66.7	35.4	58.4
Zhou & Syllables	85.7	77.9	58.3	60.0	69.3

Table 6. Quality of relevance ranking for the Portuguese corpus, including results after the syllable application; values in percentage

Method	Test "A"	Test "B"	Test "C"	Test "D"	Test "E"
Syllables isolated	73.3	65.4	60.3	69.4	66.6
Sc	56.6	48.1	48.4	47.9	49.7
Sc & Syllables	80.0	63.0	65.1	70.1	69.8
SPQ	56.7	53.1	54.0	46.5	59.1
SPQ & Syllables	73.3	65.4	68.3	70.8	71.1
Tf-Idf	56.7	61.7	59.5	68.8	65.0
Tf-Idf & Syllables	70.0	70.4	71.4	75.7	74.1
Zhou	46.7	62.7	59.5	56.3	62.0
Zhou & Syllables	80.0	69.1	69.8	72.9	72.2

Table 7. Quality of relevance ranking for the English corpus, including results after the syllable application; values in percentage

Method	Test "A"	Test "B"	Test "C"	Test "D"	Test "E"
Syllables isolated	83.8	69.3	59.5	59.2	66.9
Sc	81.1	61.4	51.4	35.5	55.0
Sc & Syllables	91.9	71.6	61.3	60.5	68.5
SPQ	64.9	61.4	50.5	36.9	55.0
SPQ & Syllables	91.9	73.9	65.8	61.9	70.4
Tf-Idf	54.1	51.1	52.3	39.5	51.8
Tf-Idf & Syllables	75.7	72.7	64.9	61.8	66.9
Zhou	51.4	52.3	52.3	42.1	56.0
Zhou & Syllables	89.2	73.9	61.3	59.2	68.5

Table 8. Quality of relevance ranking for the Spanish corpus, including results after the syllable application; values in percentage

According to the previous tables we can see that the results of almost all methods are satisfactory, with almost all results being superior to 60% (and to 80% in some cases). First of all, it should be mentioned that for the "A" test set, the one that tests the 100 more frequent words, *Tf-Idf* and *Zhou & Slater* methods are inefficient as expected. For instance, while in

table 6 (Portuguese corpus) *SPQ* has values of 71.4% of quality for this test set, *Tf-Idf* and *Zhou & Slater* methods have 46.4% and 25% respectively. Second, almost all methods (excluding syllable application) start to fail in "C" and "D" test sets. This has probably to do with the fact that those test sets are made from words with lower frequency in the corpora, because although statistical methods should be frequency independent, the frequency factor for the analysis of statistical data is always present. The situation is more serious in the "D" test set which has words with lower frequency (with words having frequencies of 2) which makes *Score* and *SPQ* methods to fail with quality results below 50%.

Comparing *Score*, *SPQ*, *Tf-Idf* and *Zhou & Slater* methods directly it can be noted that in a general way, in the "C", "D" and "E" test sets they have almost the same kind of results (despite some minor exceptions). It should be noted however that for the test set "A", the metrics *Score* and *SPQ* are more efficient than the other two because *Tf-Idf* and *Zhou & Slater* methods tend to damage frequent relevant words. Also it should be noted that *SPQ* metric is, as mentioned before, more efficient in Portuguese and Spanish languages than in English.

Considering now the syllable method, it can be noted that as an isolated metric, it has good results having almost the best results when considering the other isolated methods (without syllable application). When we consider the application of the syllable method to the other methods it can be noted that it improves greatly almost all results, including the results of *Tf-Idf* and *Zhou & Slater* methods, being the most flagrant case the rise of 25% to 85.7% of the *Zhou & Slater* method in the Portuguese corpus. Also, for the "D" test set, the most problematic one, it can be noted that in average, the quality results are above 60% for the Portuguese and Spanish corpus and above 70% for the English corpus. For the "A" test set, which *Tf-Idf* and *Zhou & Slater* methods have low results, after the application of the syllable method to those metrics, we have, in average, quality values of 82% for the Portuguese and Spanish corpus and 75% for the English one.

The following tables (tables 9 to 11) present the results of Precision and Recall for the Islands method. The test set used to create the tables was the "E" test set because it is the most complete one, including all the words of the other test sets. It should be mentioned again that the Islands method aims to extract the relevant words in a *Boolean* basis, either by considering a word true or false, from the relevance rankings previously obtained by the other methods.

Method	Precision	Recall
Syllables isolated	76.4	78.1
Sc	70.6	85.8
Sc & Syllables	77.0	75.3
SPQ	75.6	64.9
SPQ & Syllables	82.0	72.1
Tf-Idf	80.0	59.5
Tf-Idf & Syllables	83.5	65.8
Zhou	70.1	79.1
Zhou & Syllables	78.9	77.4

Table 9. Precision and Recall values for the Islands method for the Portuguese corpus, including results after the syllable application; values in percentage

Method	Precision	Recall
Syllables isolated	68.2	82.2
Sc	61.1	76.8
Sc & Syllables	69.2	77.0
SPQ	63.6	48.4
SPQ & Syllables	71.6	65.7
Tf-Idf	73.6	47.1
Tf-Idf & Syllables	81.5	55.4
Zhou	66.7	75.4
Zhou & Syllables	71.5	76.8

Table 10. Precision and Recall values for the Islands method for the English corpus, including results after the syllable application; values in percentage

Method	Precision	Recall
Syllables isolated	73.5	78.2
Sc	68.3	84.4
Sc & Syllables	74.7	77.1
SPQ	70.9	65.4
SPQ & Syllables	78.5	70.2
Tf-Idf	72.5	46.6
Tf-Idf & Syllables	78.2	60.8
Zhou	66.5	75.9
Zhou & Syllables	76.6	75.9

Table 11. Precision and Recall values for the Islands method for the Spanish corpus, including results after the syllable application; values in percentage

According to the previous tables it can be noted that almost all the methods have good values of Precision and Recall which means that the Islands criterion is, in a certain way, resistant to the variations of each metric used (to create the relevance ranks). In the English case (table 10) it can be noted a situation previously stated: although *Tf-Idf* has a good result on Precision, it has, however, a low value of Recall. In this case it means that although the metric is considering as relevant words with an efficiency of 73.6%, it is only considering relevant about 47.1% of all the truly relevant words. This is due, however, not to the Islands method, but to the metric used (*tf-Idf*), otherwise all the values of Recall would be as low.

For the relevance rankings with the syllable method applied it can be seen (as in the previous tables 6 to 8) that the syllable method isolated can serve as a good relevance ranking metric to use in the Islands method, having average values of 70% for Precision and almost 80% for Recall. Also, in almost all cases Precision values rises with the application of the syllable methods to those metrics, breaking the frontier of 80% for the Portuguese corpus (and *Tf-Idf* in the English one), and reaching almost 80% in the Spanish corpus. About Recall, it changes, rising sometimes and lowering other times, but in average at about 75% in Portuguese and Spanish corpora, and slightly lower in the English corpus. In general, the

syllable method is able to improve the results of the Islands method as well as the quality of the relevance rankings.

6. Conclusions

The process of extraction of relevant unigrams and *n*-grams is an area of great applicability. The most flagrant examples are associated, somehow, with the classification of documents. For instance, current search engines would benefit from having unigram and multiword extractors instead of returning results merely based on the occurrence of terms as they do nowadays. Also, applications like grouping and indexing of documents are also great candidates to benefit from this kind of extractors.

However, the extraction of unigrams has been an almost ignored area by the scientific community. As it was mentioned before, to leave out unigrams in a process of extraction impoverishes the final results. The few approaches existent today suffer, however, a few problems. Essentially, they harm severely the frequent relevant words, when they are, as seen, pretty descriptive of the general topics of texts. On the other hand, all existent approaches are only capable of creating relevance rankings, which may be restrictive for some kind of applications like the characterization of keywords of documents.

In this chapter we have presented two new metrics to evaluate words that are at the same time, language, frequency and context independent. *Score* measure is capable of improving results for very frequent words, while *SPQ*, besides that, has good results for Portuguese and Spanish (or other latin-descendent languages) documents.

About the unigram extractor also presented in this chapter (Islands method), it allows to extract, with good results of efficiency, relevant unigrams from the relevance rankings. By the fact that any relevance rank can be used, this method is then metric independent.

At last, we've presented the syllable method that can work as well as an isolated metric or with another metric. It has been seen that its results are encouraging.

Although we have encouraging results, there can be, however, some improvements or further research following the sequence of this work. There is an interest in increasing even more the efficiency of all the methods, also increasing the values of Precision and Recall of the Islands method, arrange mechanisms to associate singular and plural terms and using synonyms, and, mostly, proceed with further research in the syllable area, a very promising area.

7. References

Afrin, Taniza. (2001). Extraction of Basic Noun Phrases from Natural Language using Statistical Context-Free Grammar. *Master's Thesis*. Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Das, A.; Marko, M.; Probst, A.; Porte, M. A. & Gershenson, C. (2002). Neural Net Model for featured word extraction.

Feldman, R.; Rosenfeld, B. & Fresko, M. (2006). TEG – An hybrid approach to information extraction., In *Knowledge and Information Systems*., Vol. 9 (1), pp. 1-18, Springer-Verlag, 0219-1377, New York, USA.

Gao, Y. & Zhao, G. (2005). *Lecture Notes in Computer Science*, Knowledge-Based Information Extraction: A case study of recognizing emails of Nigerian frauds., pp. 161-172, Springer Berlin, 978-3-540-26031-8, Heidelberg.

Heid, Ulrich. (2000). A Linguistic Bootstrapping Approach to the Extraction of Term Candidates from German Texts., Terminology, pp.161-181.

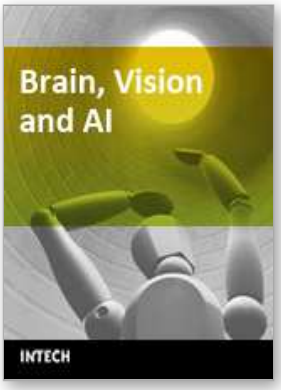
Luhn, H.P. (1958). The Automatic Creation of Literature Abstracts., *IBM Journal of Research and Development*, 2. pp. 159-165.

Ortuño, M.; Carpena, P.; Bernaola-Galván, P.; Muñoz, E. & Somoza, A.M., (2002). Keyword detection in natural languages and DNA., *Europhys.* Lett 57, pp. 759-764.

Salton, G. & McGill, M.J. (1987). Term weighting approaches in automatic text retrieval., In *Information Processing & Management*., Vol. 24 (5), pp. 513-523, 0306-4573, Pergamon Press.

Ventura, J. & Silva, J. (2007). New Techniques for Relevant Word Ranking and Extraction. In Proceedings of *13th Portuguese Conference on Artificial Intelligence*, pp. 691-702, Springer-Verlag.

Zhou, H. & Slater, G. (2003). A Metric to Search for Relevant Words., *Physica A: Statistical Mechanics and its Applications*., Vol. 329 (1), pp. 309-327.



Brain, Vision and AI

Edited by Cesare Rossi

ISBN 978-953-7619-04-6
Hard cover, 284 pages
Publisher InTech
Published online 01, August, 2008
Published in print edition August, 2008

The aim of this book is to provide new ideas, original results and practical experiences regarding service robotics. This book provides only a small example of this research activity, but it covers a great deal of what has been done in the field recently. Furthermore, it works as a valuable resource for researchers interested in this field.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Joao Ventura and Joaquim Ferreira da Silva (2008). Ranking and Extraction of Relevant Single Words in Text, Brain, Vision and AI, Cesare Rossi (Ed.), ISBN: 978-953-7619-04-6, InTech, Available from: http://www.intechopen.com/books/brain_vision_and_ai/ranking_and_extraction_of_relevant_single_words_in_text

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821