Towards Automatic Building of Document Keywords

Joaquim Silva
CITI/DI/FCT
Universidade Nova de Lisboa
jfs@di.fct.unl.pt

Gabriel Lopes
CITI/DI/FCT
Universidade Nova de Lisboa
gpl@di.fct.unl.pt

Abstract

Document keywords are associated to documents as summarized versions of the documents' content. Considering that the number of documents is quickly growing every day, the availability of these keywords is very important. Although, usually keywords are manually written. This motivated us to work on an approach to change this manual procedure for an automatic one.

This paper presents a language independent approach that extracts the most relevant Multiword Expressions and single words from documents and propose them to describe the core content of each document.

1 Introduction

Keywords provide efficient and sharp access to documents concerning their main topics, that is, their core content. Keywords are semantically relevant terms, usually being relevant noun-phrases rather than long full phrases. Full phrases such as ”John F Kennedy's speechwriter hails Obama's address” can be extracted by summarization approaches, but it wouldn't be appropriate if used as keywords since it doesn't mean any main topic/subtopic. On the other hand, by using LocalMaxs algorithm (Silva and Lopes, 1999) it is possible to extract Multiword Expressions (MWEs) from documents and, some of the most relevant ones relatively to each document can be used as that document's descriptor, if properly selected. In this paper we will show that MWEs having 2, 3 our 4 words, that is, (2-4)-gram MWEs, are the most appropriate ones to fit the typical keywords' semantic sharpness, as would be the case of “climate change”, “American Red Cross”, “social and economic policy”, etc., rather than (5-7)-grams and larger MWEs addressing more specific meanings, such as ”skills for lifelong learning process report” or ”Assessment of the use of Magnetic Resonans Tomography”.

On the other hand, although MWEs extracted by LocalMaxs algorithm are usually relevant, some of them are semantically vague or simply not relevant, such as ”general use” or ”Annex I”, not having the semantic relevance and sharpness required to form keywords. Other MWEs such as “in case of” or ”as soon as possible” may be useful for lexicon enrichment to improve Natural Language Processing, but they are not relevant MWEs to be taken as keywords.

During our investigation, we discovered that the median of the words' length in each MWE has a strong influence in the MWE relevance. Thus, combining this and other factors that influence relevance, a metric, Mk, is proposed to better evaluate the relevance of each MWE under the purpose of obtaining keywords, and consequently its relevance score in each document.

Although most document keywords are multiwords, there are some single words , that is, 1-grams, whose strong and sharp meaning make them good keywords, such as ”Agriculture”, ”salmonella”, among others. Then, since we wanted to include single words in the set of the main keywords of each document, and because LocalMaxs algorithm does not extracts 1-grams, we had to select the most informative single words

Coling 2010: Poster Volume, pages 1149–1157,
Beijing, August 2010

from documents using another metric, Sk, also presented in this paper.

This paper proposes a statistical and language-independent approach to generate document descriptors based on the automatic extraction of the most informative MWEs and single words, in terms of document summarization, under the purpose of keywords, taken from each document. Next section analyzes related work. A brief explanation of the LocalMaxs algorithm is presented in section 3. In section 4 we propose the metrics Mk and Sk and consider other measures. Results are presented in section 5 and conclusion are made in the last section.

2 Related Work

In (Cigarrán et al., 2005; Liu et al., 2009; Hulth, 2004) authors propose extraction of noun phases and keywords. However, these are not language-independent approaches, since they use some language-dependent tools such as stop-words removing, lemmatization, part-of-speech tagging or syntactic pattern recognition.

In (Delort et al., 2003), authors address the issue of Web document summarization by context. They consider the context of a Web document by the textual content of all documents linking to it. According to the authors, the efficiency of this approach depends on the size of the content and context of the target document. However, its efficiency also depends on the existence of links to the target documents.

In (Aliguliyev, 2006) a generic summarization method is proposed. It extracts the most relevance sentences from the source document to form a summary. The summary can contain the main contents of different topics. This approach is based on clustering of sentences and, although results are not shown, it does not use language-dependent tools.

Other Information Extraction methods rely on predefined linguistic rules and templates to identify certain entities in text documents (Yangarber and Grishman, 2000; Jacquemin, 2001). Again, these are not language-independent approaches, despite the good results that they give rise to.

Some approaches address specific-domain problems. In (Alani et al., 2003), authors propose a method to extract artist information, such as name and date of birth from documents and then generate his or her biography. It works with meta-data triples such as (subject-relation-object), using ontology-relation declarations and lexical information. Clearly, this approach is not language-independent. In (Velardi et al., 2001), a method to extract a domain terminology from available documents such as the Web pages is proposed. The method is based on two measures: Domain Relevance and Domain Consensus that give the specificity of a terminological candidate. In (Martínez-Fernández et al., 2004) the News specific-domain is addressed. Again, this approach is not language-independent.

A supervised approach (Ercan and Cicekli, 2007) extracts keywords by using lexical chains built from the WordNet ontology (Miller, 1991), a tool which is not available for every language.

Rather than being dependent on specific languages, structured data or domain, we try to find out more general and language-independent features from free text data.

In (Silva and Lopes, 2009), a MWEs extractor and a metric, $LeastRvar$, extracts keywords from documents. However, single words are ignored as possible keywords and their global results are outperformed by our proposal.

3 Using LocalMaxs Algorithm to Extract Keyword Candidates

We used the SCP_f cohesion metric and the LocalMaxs algorithm to extract MWEs from document *corpora*. Although details about these tools are given in (Silva and Lopes, 1999; Silva et al., 1999), here follows a brief description for paper self-containment. Thus, LocalMaxs is based on the idea that each n-gram[1] has a kind of *glue* or cohesion sticking the words together within the n-gram. Different n-grams usually have different cohesion values. One can intuitively accept that there is a strong cohesion within the n-gram ”Giscard d’Estaing” i.e. between the words ”Giscard” and ”d’Estaing”. However, one cannot say that there is a strong cohesion within the 2-grams ”or given” or within the ”of two”. Thus, in order to

[1] $w_1 \dots w_n$ or $(w_1 \dots w_n)$ are also used to denote an n-gram of length n.

measure the cohesion value not only of 2-grams, but also for every n-gram of any size in the corpus, we used the $SCP\_f(.)$ metric:

$$SCP\_f(w_1 \dots w_n) = \frac{p(w_1 \dots w_n)^2}{Avp} \qquad (1)$$

$$Avp = \frac{1}{n-1} \sum_{i=1}^{n-1} p(w_1 \dots w_i) \, . \, p(w_{i+1} \dots w_n) \qquad (2)$$

where $p(w_1 \dots w_n)$ is the probability of the n-gram $w_1 \dots w_n$ in the corpus. This way, any size n-gram is *transformed* in a pseudo-bigram that reflects the *average cohesion* between any two adjacent contiguous sub-n-gram of the original n-gram. Now it is possible to compare cohesions from n-grams of different sizes.

3.1 LocalMaxs Algorithm

LocalMaxs is a language independent algorithm to filter out cohesive n-grams of text elements (words, tags or characters), requiring no threshold arbitrarily assigned.

Definition 1. *Let $W = w_1 \dots w_n$ be an n-gram and $g(.)$ a cohesion generic function. And let: $\Omega_{n-1}(W)$ be the set of $g(.)$ values for all contiguous $(n-1)$-grams contained in the n-gram W; $\Omega_{n+1}(W)$ be the set of $g(.)$ values for all contiguous $(n+1)$-grams which contain the n-gram W, and let $len(W)$ be the length (number of elements) of n-gram W. So, it is stated that*

W is a Multi Element Unit (MEU) if and only if,
for $\forall x \in \Omega_{n-1}(W), \forall y \in \Omega_{n+1}(W)$

$$(len(W) = 2 \wedge g(W) > y) \quad \vee$$
$$(len(W) > 2 \wedge g(W) > \frac{x+y}{2}) \ .$$

Then, for n-grams with $n \geq 3$, LocalMaxs algorithm elects every n-gram whose cohesion value is greater than the average of two maxima: the greatest cohesion value found in the contiguous $(n-1)$-grams contained in the n-gram, and the greatest cohesion found in the contiguous $(n+1)$-grams containing the n-gram. Thus, in the present approach we used LocalMaxs as a MWEs extractor — MWEs are MEUs where the elements are words — and used $SCP\_f(.)$ cohesion measure as the $g(.)$ function referred in the algorithm definition above.

4 Selecting Keywords from MWEs

Not every MWE extracted by LocalMaxs has equal relevance or semantic sharpness. Some MWEs are vague in terms of semantic sharpness, such as “important meeting” or “general use”; other ones are very specific in terms of the topic they point to, for example “Assessment of the use of Magnetic Resonans Tomografy”; some others are (2-4)-gram strongly informative MWEs, fitting the semantic sharpness of typical keywords such as “computer science” or ”Food and Agriculture Organization”, and will be privileged by the metric we present in subsection 4.4.

Some single words have adequate semantic sharpness to be included as keywords, such as ”Algebra” or ”Agriculture”, among others. However, most single words are not informative enough for that purpose.

As a consequence, we felt the need to work on adequate metrics to value and privilege the strongly informative MWEs and single words in order to find keywords in documents.

4.1 The *Tf-Idf* Metric

$Tf{-}Idf$ (Term frequency−Inverse document frequency) is a statistical metric often used in IR and text mining. Usually, it is used to evaluate how important a word is to a document in a *corpus*. The importance increases proportionally to the number of times a word/multiword appears in the document but it is offset by its frequency in the *corpus*. Thus, this is one of the metrics with which we will try to privilege the most informative MWEs and 1-grams in each document.

$$Tf{-}Idf(W, d_j) = p(W, d_j) \, . \, Idf(W, d_j) \qquad (3)$$

$$p(W, d_j) = \frac{f(W, d_j)}{N_{d_j}} \qquad (4)$$

$$Idf(W, d_j) = \log \frac{\|\mathcal{D}\|}{\|\{d_j : W \in d_j\}\|} \qquad (5)$$

where $f(W, d_j)$ if the frequency of word/multiword W in document d_j and N_{d_j} stands for the number of words of d_j; $\|\mathcal{D}\|$ is the number of documents of the *corpus*. So, $Tf{-}Idf(W, d_j)$ will give a measure of the importance of W, that is a MWE or a single word, within the particular document d_j. By the structure of term Idf we can see that it privileges MWEs and single words occurring in less documents, particularly those occurring in just one document.

4.2 The *LeastRvar* Metric

Most weakly relevant MWE and errors extracted by LocalMaxs begin or end with a so called stop-word, that is a highly frequent word appearing in most documents. However, stop-words may exist in the middle of a relevant MWE, for example "United States of America" or "Life on Mars"; but usually not in the leftmost or rightmost word of the MWEs. By considering this, $LeastRvar$ was proposed in (Silva and Lopes, 2009):

$$LeastRvar(MWE_i) = least(Lrv, Rrv) \qquad (6)$$

where $Lrv = Rvar(leftmostw(MWE_i))$,
$Rrv = Rvar(rightmostw(MWE_i))$

and

$$Rvar(W) = \frac{1}{\|\mathcal{D}\|} \sum_{d_i \in \mathcal{D}} \Big(\frac{p(W, d_i) - p(W, .)}{p(W, .)} \Big)^2 \ . \qquad (7)$$

$p(W, .)$ means the average probability of the word W considering all documents. $Rvar(.)$ is applied to the leftmost and the rightmost word of the MWE:

$$p(W, .) = \frac{1}{\|\mathcal{D}\|} \sum_{d_i \in \mathcal{D}} p(W, d_i). \qquad (8)$$

$Rvar(W)$ measures the variation of the probability of the word W along all documents. Apparently the usual formula of the variance (the second moment about the mean), would measure that variation; however, it would wrongly benefit the very frequent words such as "of", "the" or "and", among others. This happens because the absolute differences between the occurrence probabilities of any of those frequent words along all documents is high, regardless of the fact that they usually occur in every document. These differences are captured and over-valued by the variance since it measures the average value of the quantity $(distance\, from\, mean)^2$, ignoring the *order of magnitude* of the individual probabilities. Then, $Rvar(.)$ divides each *individual distance*, in the original formula of the variance, by the order of magnitude of these probabilities, that is, the mean probability, given by $p(W, .)$; see equations 7 and 8.

Then, $LeastRvar(MWE_i)$ is given by the least $Rvar(.)$ values considering the leftmost word and the rightmost word of MWE_i. This way, $LeastRvar(.)$ tends to privilege informative MWEs and penalize those multiword expressions having semantically meaningless words in the begin or in the end of it.

4.3 The *LeastCv* metric

In oder to try to obtain better results than those produced by *LeastRvar*, we changed $Rvar(.)$ to an alternative to measure the relative variation of the probability of the leftmost and rightmost words in MWEs. Then we defined:

$$LeastCv(MWE_i) = least(Lcv, Rcv) \qquad (9)$$

where $Lcv = Cv(leftmostw(MWE_i))$,
$Rcv = Cv(rightmostw(MWE_i))$,

$$Cv(W) = \sigma(W)/\mu(W) \ , \qquad (10)$$

$$\sigma(W) = \sqrt{\frac{1}{\|\mathcal{D}\|} \sum_{d_i \in \mathcal{D}} (p(W, d_i) - p(W, .))^2} \ , \qquad (11)$$

and

$$\mu(W) = p(W, .) \ ; \qquad (12)$$

$p(W, d_i)$ and $p(W, .)$ have the same meaning as in equation 7. The reader may recognize $Cv(.)$ as the *coefficient of variation*, which is given by the ratio of the standard deviation σ to the mean μ. Results in section 5 will show that $LeastCv$ also tends to privilege informative MWEs.

4.4 Two New Metrics to Find Keywords

Considering the results obtained for *LeastRvar* and *LeastCv*, as we will see in section 5, we wanted to develop a better metric to find MWE keywords and another one for single word keywords. They were built by combining some important factors that we present next.

The Median of the MWE Words' Length: Since most of the semantically meaningless words are small and long words usually have sharp meaning, we considered the median length of the words in each MWE to help on selecting the most informative MWEs. By comparison, median length showed better results than average length. For example, MWE ”Language Institute” has an average word length of 8.5 characters, but the semantically equivalent ”Institute of Languages” has a different average length of 6.66. On the contrary, the median length for both MWEs presents more close values: ($(8+9)/2 = 8.5$) for ”Language Institute” and 9 for ”Institute of languages” (the middle number after sorting the MWE words length: 2, 9 and 9). Thus, because the median values is more robust to outliers than the average value, the length of the meaningless word ”of” was, say, *ignored* in the median calculation. In fact, those equivalent meaning MWEs have similar median length values (8.5 and 9), but not so similar average length values (8.5 and 6.66). Furthermore, the robustness of the median length enables more similar values when considering MWEs in English and other equivalent MWEs in other languages where stop words are more used; for example ”écoles de conduite” (driving schools), ”producción de batata” (potato production), etc..

How Many Words for a Keyword? As the reader may check in documents having associated keywords, we noticed that the main document keywords are usually (2-4)-grams. So, we defined a factor, $Ckl(MWE_i)$, to measure how similar is the *pseudo number of words* of MWE_i to the *typical* number of words of keywords. We define the *pseudo number of words* of a MWE:

$$Pnw(MWE_i) = \frac{NumChars(MWE_i)}{Med(MWE_i)} \quad . \tag{13}$$

$NumChars(MWE_i)$ stands for the number of characters of MWE_i and $Med(MWE_i)$ is the median length of its words. $Pnw(MWE_i)$ gives a value close to the number of meaningful words of MWE_i. For example, Pnw(”Institute of Languages”) $= 20/9 = 2.2$ (close to 2); Pnw(”European Council”) $= 15/7.5 = 2$, etc.. Now, $Ckl(.)$ is given by:

$$Ckl(MWE_i) = \frac{1}{|Pnw(MWE_i) - T| + 1} \ , \tag{14}$$

where T is the *typical* number of words of the keywords. Maximum value for $CkLen(MWE_i)$ is 1; it happens if $Pnw(MWE_i)$ equals to T. As we will see by the results in section 5, we tried two T values: 2.5 and 3.5; and compared results.

The *Mk* Metric for MWE Keywords: We built $Mk(.)$ metric by improving $LeastRvar(.)$:

$$Mk(M) = LeastRvar(M).Med(M).Ckl(M) \tag{15}$$

Thus, $Mk(.)$ privileges MWEs having not only informative leftmost and rightmost words, but also having long words and a *pseudo number of words* close to the number of words of typical keywords – for reasons of lack of space, we used M instead of MWE_i in equation 15 –.

The *Sk* Metric for *Single Word* Keywords: We built $Sk(.)$ from $Rvar(.)$ – see equation 7 – to measure how meaningful is each single word:

$$Sk(W_i) = Rvar(W_i).Len(W_i) \ . \tag{16}$$

$Len(W_i)$ means the length of words W_i. Thus, $Sk(.)$ privileges single words having, not only a high relative variation of their probabilities along all documents, but also being long words.

5 Results

We analyze the quality of the document descriptors after applying the LocalMaxs extractor followed by each of the six different metrics to three different document *corpora*, each one for a different language: English, French and Spanish. Metrics applied to MWEs were $Tf{-}Idf$, $LeastCv$, $LeastRvar$, Mk [2.5] – that is $T = 2.5$ in equation 14; and Mk [3.5]. Metrics applied to single words were $Tf{-}Idf$ and Sk.

5.1 The Document Descriptor

We decided to represent the core content of each document by using its 15 most informative terms, in the sense of keywords: 11 MWEs and 4 single words. An independent evaluation criteria were defined by Prof. Francisca Xavier from the Linguistics Department of *Universidade Nova de Lisboa*. It was considered that, for example, “aim of mission” and “16 December 2003” are wrong keywords, as the first one is a too vague noun phrase and the second one, just a simple date. Relevant MWEs such as “nuclear weapons” and “financial crisis” were evaluated as keywords. However, although some proposed multi-word expressions are not keywords, they are informative in the context of the descriptor and correspond to well formed morphosyntactic tags, for example, ”56% of GDP” or ”comfort zone”: these *near-miss* cases were classified as half-correct half-wrong terms; the same classification was given to single words such as ”macro-economic” – see table 7 – which, although it’s not a noun, it’s an informative adjective.

Thus, for each document, the extracted MWEs are sorted according to each metric and the top 11 MWEs are taken as the document’s MWEs descriptor. The single words of the document are also sorted according to one of the two applied metrics ($Tf{-}Idf$ or Sk). By ignoring the rest of the MWEs and single words, there is document information which will be *lost* by these descriptors, but they must be taken as core content descriptors, not as complete/detailed reports of the documents. Although descriptors are composed by MWEs and single words, for better comparison of the metrics, tables separately show MWE descriptors or *single word* descriptors. Table 1 shows an example of a document MWE descriptor resulting from the application of one of the metrics (Mk) to the document’s MWEs extracted by LocalMaxs algorithm:

Table 1: Example of an English Document MWE Descriptor – Application of the $Mk\,[2.5]$ Metric.

enterprise profits
comfort zone
medium-sized enterprises
brain drain
cold war
Balance of Payment
56% of GDP
excessive deficit
looking ahead
exports and imports
Stability and Growth Pact

5.2 The Multi-Language *Corpora* Test

We used the *EUR-Lex corpora*, http://eur-lex.europa.eu/en/, containing European Union law documents about several topics in several European languages. We took 60 documents written in each language, English, French and Spanish to form three different *sub-corpora*. These are unstructured row text documents.

To evaluate the approach’s performance, we used Precision and Recall concepts. Precision was given by the number of keywords in the set of the 11 most scored MWEs proposed as descriptor, by the combination LocalMaxs$-$metric used, divided by 11. Recall was given by the number of keywords that are simultaneously in the document’s descriptor proposed and in the set made of the 11 most informative keywords of the document, divided by 11.

According to the criteria mentioned above, this is the evaluation of the descriptor shown in table 1, considering Precision: 8 MWEs can be accepted as keywords (1st, 3rd, 4th, 5th, 6th, 8th, 10th and 11th); 2 near-miss MWEs (2nd and 7th); and 1 weak or wrong MWE (9th). So, precision is $(8 + 2 * 0.5)/11 = 0.818$. Concerning the document this descriptor represents, there are 3 strong keywords that should be in the descriptor, but they weren’t: ”financial crisis”, ”structural reforms” and ”macroeconomic imbalances”. Thus, Recall is $8/11 = 72.7$ for this case.

5.3 Results for Different Metrics and Languages

By table 2 we may see that for the same metric, Precision or Recall values are similar for English, French and Spanish. So, this approach does not seem to privilege any of these languages, and we believe that probably this happens for many other languages, as no specific morphosyntactic information was used. Even the difference between Recall values for Spanish and English produced by $LeastRvar$ (0.61 and 0.63) would probably decrease if the test *corpora* had more documents. Table 2 also shows that $Tf{-}Idf$ presents the poor-

Table 2: Precision and Recall Average Values for the Document MWE Descriptors.

Language	Metric	Precision	Recall
English	Tf–Idf	0.51	0.35
	$LeastCv$	0.62	0.61
	$LeastRvar$	0.65	0.63
	$Mk\,[2.5]$	**0.76**	**0.72**
	$Mk\,[3.5]$	0.74	0.68
French	Tf–Idf	0.50	0.35
	$LeastCv$	0.62	0.60
	$LeastRvar$	0.64	0.63
	$Mk\,[2.5]$	**0.75**	**0.71**
	$Mk\,[3.5]$	0.73	0.68
Spanish	Tf–Idf	0.51	0.34
	$LeastCv$	0.61	0.60
	$LeastRvar$	0.64	0.61
	$Mk\,[2.5]$	**0.75**	**0.72**
	$Mk\,[3.5]$	0.74	0.67

Table 3: Example of an English Document MWE Descriptor – Application of the Tf–Idf Metric.

in the new Member States
in the new Member
new Members
Single Market
income convergence
some of the new Member
financial crisis
structural reforms
new and old
euro area
reap the full benefits of the Single Market

Table 4: Example of an English Document MWE Descriptor – Application of the $LeastRvar$ Metric.

five years
Cold War
old Members
enterprise profits
Central Bank
Excessive Deficit
medium-sized enterprises
comfort zone
56% of GDP
1.5% of GDP
brain drain

est results. In fact, due to its structure — see equation 3 — we can see that MWEs that occur many times in just one document are the most valued/privileged ones. This explains why the descriptors made by this measure tend to include too specific/local MWEs, regardless of some important ones. Table 3 shows a document descriptor generated by the combination LocalMaxs–Tf–Idf: for example MWE ”new Members” occurs in just one document, 10 times; however, ”new Members” is not a keyword. This is the descriptor of the same document from where other descriptors were generated by the combinations including $LeastRvar$ and $Mk\,[2.5]$, and shown in tables 4 and 1.

For reasons of space limitation we don’t show descriptors produced by $LeastCv$ and $MK\,[3.5]$ metrics. However, table 2 shows that $LeastCv$ was outperformed by $LeastRvar$. This table also shows that $Mk\,[2.5]$ metric presents the highest Precision (0.76, 0.75 and 0.75 for English, French and Spanish). The highest Recall values are also obtained for the same metric: 0.72, 0.71 and 0.72 for the same languages.

Tables 5 and 6 show examples of MWE descriptors of French and Spanish documents, by the application of $Mk\,[2.5]$ as it produced the best results.

Tables 7 and 8 show examples of *single word* descriptors for the same document described in table 1. As we could expect, Precision and Recall values for *single word* descriptors are lower than the values for MWEs descriptors, since singles words are usually semantically less sharp than multiwords: see table 9. Sk shows better performance than Tf–Idf, specially for Recall.

6 Conclusions

Keywords are semantic tags associated to documents, usually declared manually by users. These tags form small document descriptors and enable applications to access to the summarized documents’ core content. This paper proposes an approach to automatically generate document de-

Table 5: Example of a French Document MWE Descriptor – Application of the $Mk\,[2.5]$ Metric.

moto-fraises et motofaucheuses
agrumeraies et oliveraies
hommes Travail
Fumier liquide
familiale occupée
Mieux légiférer
d'arbres fruitiers
Superficie irriguée
Main-d'oeuvre non familiale
activités lucratives
Alignements d'arbres

Table 6: Example of a Spanish Document MWE Descriptor – Application of the $Mk\,[2.5]$ Metric.

ingredientes de cosméticos
combinaciones de ingredientes
someter a ensayo
Sustancias y Preparados
toxicidad aguda
irritación ocular
fototoxicidad aguda
explicaciones dadas
corrosión cutánea
animales utilizados
Sustancias y Preparados Químicos

Table 7: Example of an English Document *Single Word* Descriptor – Application of the Sk Metric.

vulnerabilities
growth-enhancing
post-enlargement
macro-economic

Table 8: Example of an English Document *Single Word* Descriptor – Application of the $Tf\text{–}Idf$ Metric.

economic
new
enlargement
reforms

Table 9: Precision and Recall Average Values for the Document *Single Word* Descriptors.

Language	Metric	Precision	Recall
English	$Tf\text{–}Idf$	0.52	0.36
	Sk	**0.55**	**0.48**
French	$Tf\text{–}Idf$	0.51	0.37
	Sk	**0.54**	**0.47**
Spanish	$Tf\text{–}Idf$	0.52	0.37
	Sk	**0.56**	**0.48**

scriptors, as a language-independent and domain-independent alternative to related work from other authors. This approach uses LocalMaxs algorithm to extract MWEs, and two new statistical metrics, Mk and Sk, to select the 15 most relevant MWEs and single words from each document in order to form document descriptors.

Comparing the results produced by Mk with the second best metric, $LeastRvar$, we may conclude that the introduction of the median of the words' length of each MWE and the preference for (2-4)-grams, improve the quality of document descriptors by about 11% and 9% for Precision and Recall, respectively. Furthermore, by comparison of $Mk\,[2.5]$ and $Mk\,[3.5]$ results we conclude that keywords are mostly (2-3)-grams, rather than (3-4)-grams or longer n-grams.

Results also showed that Precision and Recall values are similar for the three languages tested (English, French and Spanish), which enable us to expect similar performance to other languages. Apart from the Precision and Recall values, document descriptors made by this approach does indeed capture the core content of each document. We believe this may contribute to improve document summarization. Future work will include tests in other languages and we will work to improve results, specially for single words.

References

Alani, Harith, Kim Sanghee, David E. Millard, Mark J. Weal, Paul H. Lewis, Wendy Hall and Nigel Shadbolt. 2003. Automatic Extraction of Knowledge from Web Documents. In *Proceedings of Workshop of Human Language Technology for the Semantic Web and Web Services, 2nd International Seman-*

tic Web Conference. October 20th, Sanibel Island, Florida, USA.

Aliguliyev, Ramiz M. 2006. A Novel Partitioning-Based Clustering Method and Generic Document Summarization. In *Proceedings of the 2006 IEEE/Web Intelligence/Association for Computing Machinery and the Intelligent Agent Technology International Conference (2006 Workshops)(WI-IATW'06)*. December 18-22, Hong Kong, China.

Cigarrán, Joan. M., Anselmo Peas, Julio Gonzalo and Felisa Verdejo. 2005. Automatic Selection of Noun Phrases as Document Descriptors in an FCA-Based Information Retrieval System. B. Ganter and R. Godin (Eds.). ICFCA 2005, *Lecture Notes in Computer Science* 3403, pp. 49-63. Springer-Verlag.

Ciravegna, Fabio, Alexeie Dingli, David Guthrie and Yorick Wilks. 2003. Mining Web Sites Using Unsupervised Adaptive Information Extraction. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*. April, 12-17. Budapest, Hungary.

Delort, J.-Y., B. Bouchon-Meunier and M. Rifqi. 2003. Enhanced Web Document Summarization Using Hyperlinks. In *Proceedings of the fourteenth Association for Computing Machinery conference on Hypertext and hypermedia*. August 26-30, Nottingham, UK.

Ercan, Gonenc and Ilyas Cicekli. 2007. Using lexical chains for keyword extraction. *Information Processing and Management: an International Journal archive*. Volume 43, Issue 6, November, Pages 1705-1714, Pergamon Press, Inc.. ISSN 0306-4573.

Hulth, Anette. 2004. Enhancing linguistically oriented automatic keyword extraction. *Proceedings of Human Language Technology-North American Association for Computational Linguistics 2004 conference*. Pag.17-20. May 02-07. Boston, Massachusetts. Publisher: Association for Computational Linguistics, Morristown, NJ, USA.

Jacquemin Christian. 2001. *Spotting and Discovering Terms through Natural Language Processing*. MIT Press, ISBN 0262100851.

Liu, Feifan, Deana Pennell, Fei Liu and Yang Liu. 2009. Unsupervised approaches for automatic keyword extraction using meeting transcripts. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. May 31-June 05. Boulder, Colorado.

Martínez-Fernández, J. L., A. García-Serrano, P. Martínez, J. Villena. 2004. Automatic Keyword Extraction for News Finder. *Lectures Notes in Artificial Intelligence*, Springer-Verlag, volume 3094, pages 99–119.

Miller, George A. 1991. *The science of words.* Scientific American Library, New York.

Silva, Joaquim and Gabriel Lopes. 1999. A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multi-word Units. *In Proceedings of the 6th Meeting on the Mathematics of Language*, pages 369-381. 23-25 July, University of Central Florida, Orlando.

Silva, Joaquim and Gabriel Lopes. 2009. A Document Descriptor Extractor Based on Relevant Expressions. *14 Encontro Portugus para a Intelignicia Artificial (Fourteenth Portuguese Conference on Artificial Intelligence)*. October 12-15. Univerity of Aveiro. Lectures Notes in Artificial Intelligence, Springer-Verlag, volume 5816, pages 646-657.

Silva, Joaquim, Gael Dias, Sylvie Guilloré and Gabriel Lopes. 1999. Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multi-word Lexical Units. 9th Portuguese Conference on Artificial Intelligence. September, vora,Portugal. *Lectures Notes in Artificial Intelligence*, Pedro Barahora and Jos Alferes (Eds.). Springer-Verlag, volume 1695, pages 113-132.

Yangarber, Roman and Ralph Grishman. 2000. Machine Learning of Extraction Patterns from Unanotated Corpora: Position Statement. *Workshop on Machine Learning for Information Extraction. Held in conjunction with the 14th European Conference on Artificial Intelligence (ECAI).* 21 August. Berlin, Humboldt University.

Velardi, Paula, Michele Missikoff, and Roberto Basili. 2001. Identification of relevant terms to support the construction of Domain Ontogies. *Association for Computational Linguistics-European Association for Computational Linguistics Workshop on Human Language Technologies.* July 6-7. Toulouse, France.