AN EXTENSIVE COMPARISON OF METRICS FOR AUTOMATIC EXTRACTION OF KEY TERMS

Luis F. S. Teixeira[2], Gabriel P. Lopes[2] and Rita A. Ribeiro[1]
[1] *CA3-Uninova, Campus FCT/UNL, 2829-516 Caparica, Portugal*
[2] *CITI, Dep. Informática, FCT/UNL, 2829-516 Caparica, Portugal*

Keywords: Document keywords, Document topics, Words, Multi-words, Prefixes, Automatic extraction, Suffix arrays.

Abstract: In this paper we compare twenty language independent statistical-based metrics for key term extraction from any document collection. While some of those metrics are widely used, others were recently created. Two different document representations are considered in our experiments. One is based on words and multi-words and the other is based on word prefixes of fixed length (5 characters for the experiments made) for handling morphologically rich languages, namely Portuguese and Czech. English is also experimented, as a non-morphologically rich language. Results are manually evaluated and agreement between evaluators is assessed using k-Statistics. The metrics based on Tf-Idf and Phi-square proved to have higher precision and recall. The use of prefix-based representation of documents enabled a significant improvement for documents written in Portuguese.

1 INTRODUCTION

A key term, a keyword or a topic of a document is any word or multi-word (taken as a sequence of two or more words, expressing clear cut concepts) that reveals important information about the content of a document from a larger collection of documents.

Extraction of document key terms is far from being solved. However this is an important problem that deserves further attention, since most documents are still (and will continue to be) produced without explicit indication of their key terms as metadata. Moreover, most existing key term extractors are language dependent and, as such, require linguistic processing tools that are not available for the majority of the human languages. Hence our bet, namely in this paper, on language independent extraction of key terms.

So, the main aim of this work is to compare metrics for improving automatic extraction of key single and multi-words from document collections, and to contribute to the discussion on this subject matter.

Our starting point was the work by (Silva & Lopes, 2009), on multiword key term extraction, where two basic metrics were used: Tf-Idf and relative variance (Rvar). By looking more carefully at the examples shown in (Silva & Lopes, 2009), where plain Tf-Idf metric is used, it became apparent that, the comparison made between the two metrics was unfair. A fair comparison would require the use of a Tf-Idf derived metric taking into account the Tf-Idf values for multi-word extremities as well as the medium character length per word of each multi-word as it is proposed for the use of Rvar variant metric in (Silva & Lopes, 2009). Moreover, as one needs to calculate the relevance of words and multi-words using the same metrics, we decided to rank simultaneously words and multi-words describing the content of any document in a collection according to the metric assigned value to that word or multi-word. This diverges from the proposal made in (Silva & Lopes, 2010) where "a priori" fixed proportion of words is required. And no one knows "a priori" which documents are better described by words alone or by multi-words. Nor does she/he know the best proportion of key words or key multi-words.

This way, our work improves the discussion started in (Silva & Lopes, 2009), and continued in (Silva & Lopes, 2010), but we arrive at different conclusions, namely that Tf-Idf and Phi-square based metrics enabled higher precision and recall for the extraction of document key terms. The use of a prefix-based representation of documents enabled a

F. S. Teixeira L., P. Lopes G. and A. Ribeiro R..
AN EXTENSIVE COMPARISON OF METRICS FOR AUTOMATIC EXTRACTION OF KEY TERMS.
DOI: 10.5220/0003720400550063
In *Proceedings of the 4th International Conference on Agents and Artificial Intelligence* (ICAART-2012), pages 55-63
ISBN: 978-989-8425-95-9

significant improvement for documents written in Portuguese and a minor improvement for Czech, as representatives of morphologically rich languages, regarding precision results.

Additionally we also extend the preliminary discussion started in (Teixeira, Lopes, & Ribeiro, 2011) where some of the metrics used in current work were presented. To achieve our aims we compare results obtained by using four basic metrics (Tf-Idf, Phi-square, Mutual Information and Relative Variance) and derived metrics taking into account per word character median length of words and multi-words and giving specific attention to word extremities of multi-words and of words (*where l*eft *and right extremities of a word are considered to be identical to the word proper*). This led to a first experiment where we compare 12 metrics (3 variants of 4 metrics). On a second experiment, we decided to use a different document representation in terms of word prefixes of 5 characters in order to tackle morphologically rich languages. As it would be senseless to evaluate the relevance of prefixes, it became necessary to project (bubble) prefix relevance into words and into multi-words.

All the experimental results were manually evaluated and agreement between evaluators was assessed using k-Statistics.

This paper is structured as follows: related work is summarized in section 2; our system, the data and the experimental procedures used are described in section 3; the metrics used are defined in section 4; results obtained are shown in section 5; conclusions and future work are discussed in section 6.

2 RELATED WORK

In the area of document classification it is necessary to select features that later will be used for training new classifiers and for classifying new documents. This feature selection task is somehow related to the extraction of key terms addressed in this paper. In (Sebastiani, 2002), a rather complete overview of the main metrics used for feature selection for document classification and clustering is made.

As for the extraction of multi-words and collocations, since we also need to extract them, we just mention the work by (Silva & Lopes, 1999), using no linguistic knowledge, and the work by (Jacquemin, 2001), requiring linguistic knowledge.

In the area of keyword and key multi-word extraction, (Hulth, 2003), (Ngomo, 2008), (Martínez-Fernández, García-Serrano, Martínez, & Villena, 2004), (Cigarrán, Peñas, Gonzalo, & Verdejo, 2005), (Liu, Pennell, Liu, & Liu, 2009) address the extraction of keywords in English. Moreover those authors use language dependent tools (stop-words removing, lemmatization, part-of-speech tagging and syntactic pattern recognition) for extracting noun phrases. Being language independent, our approach clearly diverges from these ones. Approaches dealing with extraction of key-phrases (that are according to the authors "*short phrases that indicate the main topic of a document*") include the work of (Katja, Manos, Edgar, & Maarten de, 2009) where Tf-Idf metric is used as well as several language dependent tools. In (Mihalcea & Tarau, 2004), a graph-based ranking model for text processing is used. The authors use a 2-phase approach for the extraction task: first they select key-phrases representative of a given text; then they extract the most "important" sentences in a text to be used for summarizing document content.

In (Peter, 2000) the author tackles the problem of automatically extracting key-phrases from text as a supervised learning task. And he deals with a document as a set of phrases, which his classifier learns to identify as positive or negative examples of key-phrases.

(Lemnitzer & Monachesi, 2008) deal with eight different languages, use statistical metrics aided by linguistic processing, both to extract key phrases and keywords. Dealing also with more than one language, (Silva & Lopes, 2009) extract key multi-words using purely statistical measures. In (Silva & Lopes, 2010) statistical extraction of keywords is also tackled but a predefined ratio of keywords and key multi-words is considered per document, thus jeopardizing statistical purity.

(Matsuo & Ishizuka, 2004) present, a keyword extraction algorithm that applies to isolated documents, not in a collection. They extract frequent terms and a set of co-occurrences between each term and the frequent terms.

In summary, the approach followed in our work is unsupervised, language independent and extracts key words or multi-words solely depending on their ranking values obtained by applying the 20 metrics announced and explained bellow in section 4.

3 SYSTEM DATA AND EXPERIMENTS

Our system is made of 3 distinct modules. First module is responsible for extracting multi-words, based on (Silva, Dias, Guilloré, & Lopes, 1999) and using the extractor of (Gomes, 2009). A Suffix

Array was used (McIlroy, 2007), (Yamamoto & Church, 2001) for frequency counting of words, multi-words and prefixes. This module is also responsible for ranking, according to the metrics used, words and multi-words per document. And it allows the back office user to define the minimum word and prefix length to be considered.

Second module is a user interface designed to allow external evaluators to classify the best 25 terms ranked according to each of selected metrics. When moving from ranking classification related to one metric to the ranking produced by another metric, evaluations already made are pervasively propagated. This feature enables evaluators to reconsider at any time some of their own earlier options.

Third module is also a user interface acting as a back office application, allowing an easy interpretation of the classifications produced by the external evaluators. It also shows graphically the k-Statistics resulting from evaluations of any two evaluators.

We worked with a collection of texts, for the three languages experimented, Portuguese (pt), English (en) and Czech (cs), from European legislation (*http://eur-lex.europa.eu/[L]/legis/latest /chap16.htm, where [L]* can be replaced by any of the following language references: pt, en or cs*)*. The texts were about Science, Dissemination of information and Education and Training. Czech corpus also included texts about Culture. Apart from these texts for Czech, the remaining of the corpus documents was parallel for the three languages. So the total number of terms for these collections was: 109449 for Portuguese, 100890 for English and 120787 for Czech.

We worked with words having a minimum length of six characters (this parameter is changeable) and filtered multi-words (with words of any length) by removing those containing punctuation marks, numbers and other special symbols. As it will be seen later some additional filtering operations will be required and discussed in the conclusions section.

Evaluators were asked to evaluate the 25 best ranked terms for a selected sub-group of six of the twenty metrics for a sub-set of five randomly selected documents of a total of 28 documents for each language. The Evaluators had access to the original documents from where key-words were extracted. When document metadata contained the keywords assigned to it, evaluators had also access to this information thus helping the evaluation task. It is worth telling that when this metadata exists, it generally does not use mutatis mutantis the multi-word terms as they are used in the document. This information was not used for the extraction task performed.

Four classifications were considered in the evaluation: “good topic descriptor” (G), “near good topic descriptor” (NG), bad topic descriptor” (B), and “unknown” (U). A fifth classification, “not evaluated” (NE), was required to enable the propagation of evaluation, for those metrics that were not specifically evaluated. In Section 5 the results of the experiments are presented.

It must be stressed that the multiword extractor used is available on the web page referred in (Gomes, 2009)

4 METRICS USED

As mentioned before, we used 4 basic metrics: Tf-Idf, Phi-square, Relative Variance (Rvar) and Mutual Information (MI).

Formally, **Tf-Idf** for a term *t* in a document d_j is defined in equations (1), (2) and (3).

$$Tf - Idf\ (t, d_j) = p(t, d_j) * Idf\ (t, d_j) \quad (1)$$

$$p(t, d_j) = f(t, d_j)/Nd_j \quad (2)$$

$$Idf(t, d_j) = \log(\|D\|/\|\{d_j : t \in d_j\}\|) \quad (3)$$

Notice that, in (1), instead of using the usual term frequency factor, probability $p(t, d_j)$, defined in equation (2), is used. There, $f(t, d_j)$, denotes the frequency of term *t* in document d_j, *t* denotes a prefix, a word, or a multiword, and Nd_j refers to the number of words or n-grams of words contained in d_j. The total number of documents present in the corpus is given by$\|D\|$. The use of a probability in (1) normalizes the Tf-Idf metric, making it independent of the size of the document under consideration.

Rvar (Relative Variance) is the metric proposed by (Silva & Lopes, 2009), defined in equation (4). It does not take into account the occurrence of a given term *t* in a specific document in the corpus. It deals with the whole corpus, and thus loses the locality characteristics of term t. This locality is recovered when the best ranked terms are reassigned to the documents where they occur.

$$\mathrm{Rvar(t)} = (1/\|\mathrm{D}\|) * \sum_{\mathrm{j}} \left(\mathrm{p(t, d_j)} - \mathrm{p(w,.)/p(t,.)}\right)^2 \quad (4)$$

$p(t, d_j)$ is defined in (2) and $p(t, .)$denotes the mean probability of term *t*, taking into account all

documents in the collection. As above, we take *t* as denoting a prefix, a word, or a multiword.

Phi-Square (Everitt, 2002) is a variant of the well-known Chi-Square metric. It allows a normalization of the results obtained with Chi-Square, and is defined in equation (5), where M is the total number of terms (prefixes, words, or multi-words) present in the corpus (the sum of terms from all documents in the collection). A denotes the number of times term t occurs in document d. B stands for the number of times term $t$occurs in documents other than d, in the collection. C means the number of terms of the document d subtracted by the amount of times term t occurs in document$d$. Finally, D is the number of times that neither document d nor term t co-occur (i.e., N-A-B-C, where N denotes the total number of documents)**.**

$$\varphi^2(t,d) = \frac{\left(\frac{N*(AD-CB)^2}{(A+C)*(B+D)*(A+B)*(C+D)}\right)}{M} \quad (5)$$

Mutual Information (Manning, Raghavan, & Schütze, 2008) is a widely used metric for identifying associations between randomly selected terms. For our purposes we used equation (6) where t, d, A, B, C and N have identical meanings as above for equation (5).

$$MI(t,d) \approx \log(A*N/(A+C)*(A+B)) \quad (6)$$

In the rest of this section we will introduce derivations of the metrics presented above for dealing, on equivalent grounds, with aspects that were considered crucial in (Silva & Lopes, 2009, 2010) for extracting key terms. Those derivations will be defined on the basis of 3 operators: Least (L), Median (M) and Bubble (B). In the equations below MT stands for any of the previously presented metrics (Tf-Idf, Rvar, Phi-square or φ^2, and Mutual Information or MI), P stands for a Prefix, W for a word, $MW$for a multi-word taken as word sequence $(w_1 \ldots w_n)$.

Least Operator is inspired by the metric LeastRvar introduced in (Silva & Lopes, 2009) and coincides with that metric if it is applied to Rvar.

$$LeastRvar(MW) = \min\left(Rvar(w_1), Rvar(w_n)\right) \quad (7)$$

$LeastRvar(MW)$ is determined as the minimum of $Rvar$ applied to the leftmost and rightmost words of MW, w_1 and w_n. In order to treat all metrics on equal grounds operator "Least" will now be applied to metric *MT*, where *MT* may be any of the metrics Tf-Idf, Rvar, φ^2, and MI as depicted in equation (9), when a multiword MW is at stake. As above, Least_MT of a multiword MW will be equal to the minimum of the MT metric value for the extremity words, w_1 or w_n, in the multi-word MW. This operator was adapted to work with words alone as in equation (8), where the Least_MT for a word W is identical to the rank value obtained for that word using metric MT. Adaptation was made by assuming that the leftmost and rightmost words of a single word coincide with the word itself.

$$Least\_MT\ (W) = MT(W) \quad (8)$$

$$Least\_MT\ (MW) = min\left(MT(w_1), MT(w_n)\right) \quad (9)$$

Bubbled Operator, another problem we needed to solve was the propagation of the relevance of each Prefix (P) to words (W) having P as a prefix.

$$Bubbled\_MT\ (W) = MT(P) \quad (10)$$

$$Least\_Bubbled\_MT\ (MW) = min\left(Bubbled\_MT(w_1), Bubbled\_MT(w_n)\right) \quad (11)$$

In bubble based metrics, the rank of a prefix is assigned to the words it prefixes. Generally it is larger than the rank assigned by the corresponding metric to the word forms it prefixes. For example, the value assigned to the 5 character prefix "techn" in a text would be propagated to all words having that prefix, namely "technology", "technologies", "techniques", if they would appear in the same text.

Median Operator was added in order to better compare the effects of using an operator similar to the one proposed in (Silva & Lopes, 2010) which took into account the median character length of words in multi-words. By doing so, we got metrics defined in equations (12) and (13), where T represents a term (word or multi-word), *LM* stands for *Least_Median* operator applied to any base metric MT and *LBM* stands for *Least_Bubble_Median operator applied to metric MT*. And *Median* of a term T is the median of character lengths of words in a multi-word or of the word at stake. For example, for a multiword made of three words, of lengths 5, 2 and 6, median length is 5.

$$LM\_MT(T) = Least\_MT(T) * Median(T) \quad (12)$$

$$LBM\_MT\ (T) = Least\_Bubble\_MT(T) * Median(T) \quad (13)$$

5 RESULTS

In this section we present some of the results obtained. We will also show that Rvar and its

related metrics behave worse than the ones based on Tf-Idf and Phi Square, contradicting results presented in the work of (Silva & Lopes, 2009).

An example of the top five terms extracted from one document, ranked by the Phi-Square metric for the worked languages is shown in Table 1. This document was about scientific and technical information and documentation and ethics.

Table 1: Top terms ranked by Phi-Square metric, manually classified as Good (G), Near Good (NG) or Bad (B), for 3 languages for a document on scientific and technical information and documentation.

Portuguese	English	Czech
ciências e as novas tecnologias (G)	group on ethics (G)	skupiny pro etiku ve vědě (G)
ciências e as novas (B)	ethics (G)	nových technologiích (NG)
ética para as ciências (G)	science and new technologies (G)	etiku ve vědě (NG)
grupo europeu de ética (G)	the ege (G)	skupiny pro etiku (G)
membros do gee (G)	ethics in science (G)	vědě a nových technologiích (NG)

As the corpus used elaborated on Science, Information dissemination, education and training, for the example document the word "science" alone was naturally demoted.

It is important to notice also that documents were in many cases short. This has a direct impact on results, as the number of relevant words and multiwords is short and most of them are irrelevant in terms of document content. As a consequence precision obtained for shorter documents is lower than for longer documents as most of the times just one term describes document content. Longer documents pose not this problem.

In the previous table, some top-ranked key terms are a sub terms of others. This has some effect on the results, because they are not mutually independent. Looking more carefully we may also notice larger, more specific, multi-words that might be rather sharp descriptors of document content as would be the case of "group on ethics in science and new technologies". We will return to this discussion on section 6.

For the same document, best performing metric based on Rvar (see Table 3) LBM_Rvar just extracted "ethics" in position 20. Other extracted top terms include names of several European personalities.

In tables 2 and 3, precision values obtained for the 5, 10 and 20 best ranked key terms extracted using different metrics are shown.

Table 2: Average precision values for the 5, 10 and 20 best terms using the best metrics, and average for each threshold.

Czech			
Metric	Prec. (5)	Prec. (10)	Prec. (20)
Tf-Idf	0.90	0.86	0.66
L Tf-Idf	0.75	0.70	0.61
LM Tf-Idf	0.70	0.65	0.59
LB Tf-Idf	0.80	0.68	0.65
LBM Tf-Idf	0.65	0.68	0.66
φ2	0.70	0.70	0.61
L φ2	0.70	0.60	0.58
LM φ2	0.70	0.60	0.58
LB φ2	0.55	0.63	0.55
LBM φ2	0.55	0.65	0.59
Average	**0.72**	**0.68**	**0.61**
English			
Metric	Prec. (5)	Prec. (10)	Prec. (20)
Tf-Idf	0.84	0.74	0.67
L Tf-Idf	0.78	0.66	0.68
LM Tf-Idf	0.81	0.78	0.66
LB Tf-Idf	0.85	0.66	0.65
LBM Tf-Idf	0.82	0.69	0.62
φ2	0.84	0.78	0.68
L φ2	0.83	0.76	0.69
LM φ2	0.87	0.78	0.70
LB φ2	0.83	0.74	0.62
LBM φ2	0.80	0.74	0.65
Average	**0.83**	**0.73**	**0.66**
Portuguese			
Metric	Prec. (5)	Prec. (10)	Prec. (20)
Tf-Idf	0.69	0.70	0.66
L Tf-Idf	0.64	0.66	0.65
LM Tf-Idf	0.68	0.63	0.64
LB Tf-Idf	0.86	0.71	0.65
LBM Tf-Idf	0.83	0.70	0.68
φ2	0.73	0.73	0.62
L φ2	0.68	0.64	0.59
LM φ2	0.61	0.64	0.59
LB φ2	0.60	0.65	0.65
LBM φ2	0.62	0.61	0.62
Average	**0.70**	**0.67**	**0.63**

Regarding recall values presented in tables 4 and 5, it is necessary to say that: 1) Tf-Idf, Phi Square and derived metrics extract very similar key terms; 2) Rvar and MI, alone, are unable to extract key terms as, depending on the length of documents, the top ranked 100, 200 or more terms are equally valuated by these metrics; 3) derived metrics of Rvar and MI extract very similar rare key terms completely dissimilar from those extracted by Tf-Idf, Phi Square and derived metrics; 4) by evaluating the 25 best ranked terms by 6 metrics (Phi Square, Least Tf-Idf, Least Median Rvar, Least Median MI, Least Bubble Median Phi Square and Least Bubble Median Rvar) we obtained from 60 to 70 terms

evaluated per document.

Recall was determined on the basis of these 60 to 70 evaluated terms. So, recall values presented in tables (4) and (5) are upper bounds of real recall values. Table 2 shows results for the metrics with best precision for the three languages, all of them with results above 0.50. Notice, that for Portuguese and Czech, the average precision is similar. The best results were obtained for the top ranked 5 terms, decreasing with similar values when dealing with the top ranked 10 and 20 terms. In average, English language presents the best results.

Also from table 2 we can point out that, for Portuguese, best results were obtained with metrics *Least Bubble Tf-Idf* and *Least Bubble Median Tf-Idf*. This means that Bubble operator and prefix representation enabled precision results closer to those obtained for English.

Tf-Idf had the best results in Czech, for all thresholds. In English, *Least Median Phi Square* enabled the best results. Moreover, for the 10 best terms threshold, English has three metrics with the best results, the one already mentioned and *Least Median Tf-Idf and Phi-Square*.

Table 3: Precision values for the 5, 10 and 20 best terms using the Rvar and MI best metrics, and average for each threshold.

Czech			
Metrics	Prec. (5)	Prec. (10)	Prec. (20)
LBM Rvar	0.50	0.39	0.27
LM Rvar	0.45	0.31	0.22
LBM MI	0.40	0.40	0.26
LM MI	0.45	0.31	0.22
Average	**0.45**	**0.35**	**0.24**
English			
Metrics	Prec. (5)	Prec. (10)	Prec. (20)
LBM Rvar	0.52	0.43	0.40
LM Rvar	0.47	0.42	0.35
LBM MI	0.46	0.49	0.43
LM MI	0.47	0.42	0.34
Average	**0.48**	**0.44**	**0.38**
Portuguese			
Metrics	Prec. (5)	Prec. (10)	Prec. (20)
LBM Rvar	0.52	0.48	0.41
LM Rvar	0.46	0.36	0.35
LBM MI	0.52	0.48	0.43
LM MI	0.42	0.35	0.33
Average	**0.48**	**0.42**	**0.38**

As pointed above, Rvar and MI metrics alone were unable to discriminate the top 5, 10 or 20 best ranked terms. This probably explains the need to use the Least and Median operators proposed by (Silva & Lopes, 2009, 2010). Precision for the Rvar and MI derived metrics is shown in table 3. It shows clearly that Tf-Idf and Phi Square based metrics, in table 2, are much better than those based on Rvar and MI. They get for the best metrics, values a bit higher than 0.50, and generally all bellow 0.50 which makes the average precision for these metrics rather poor.

In terms of "Recall" (upper bounds of recall), shown in tables 4 and 5, one of our goals was to increase the Czech recall, which we believe to have accomplished. In the same line with precision, the metrics based on Tf-Idf and Phi-Square have better recall values, in table 4, than those obtained for Rvar and MI-based metrics, in table 5. We have chosen to present "recall" values for the top 20 ranked relevant terms as these values are higher than for 5 or 10 best ranked terms. Recall values obtained for Rvar and MI derived metrics (Table 5) are much lower than those obtained for Tf-Idf and Phi-Square derived metrics, as Rvar and MI derived metrics choose rare terms that may specify very specific subject matters of documents.

Tables 6 and 7 depict the agreement between evaluators, for Portuguese and English, by using Kappa statistics. It shows that for Portuguese we have higher levels of agreement for the Tf-Idf and Phi-Square based metrics. For English agreement achieved is not so high, but never the less, we consider it acceptable.

Table 4: "Recall" Values for threshold of 20 best terms for Tf-Idf and Phi Square based metrics, and average recall.

	Czech	**English**	**Portuguese**
	P(20)	P(20)	P(20)
tfidf	0.68	0.43	0.48
L Tf-Idf	0.56	0.48	0.46
LM tfidf	0.52	0.43	0.44
LB tfidf	0.60	0.38	0.37
LBM tfidf	0.54	0.35	0.40
φ2	0.50	0.44	0.48
L φ2	0.50	0.41	0.36
LM φ2	0.51	0.43	0.37
LB φ2	0.40	0.37	0.33
LBM φ2	0.43	0.41	0.35
Average	**0.54**	**0.41**	**0.40**

Table 5: Recall Values for threshold of 20 best terms for Rvar and MI based metrics, and average recall.

	Czech	**English**	**Portuguese**
	P(20)	P(20)	P(20)
LBM Rvar	0.20	0.16	0.13
LM Rvar	0.25	0.12	0.14
LBM MI	0.20	0.16	0.15
LM MI	0.25	0.11	0.13
Average	**0.23**	**0.14**	**0.14**

Table 6: Kappa statistics-based agreement between the evaluators, for Portuguese and English, for Tf-Idf and Phi-Square based metrics.

	Portuguese	English
tfIdf	0.57	0.35
LM tfidf	0.56	0.42
LB tfidf	0.67	0.38
LBM tfidf	0.64	0.40
L φ2	0.64	0.46
LM φ2	0.56	0.40
LBM φ2	0.54	0.31

Disagreement was mainly due to acceptance of some adjectives as near good descriptors by one of the evaluators, while the other systematically rejected them in the sense that adjectives, by themselves, are not good descriptors. This means that, if the evaluation phase had been preceded by identification of a couple of cases where the evaluation would be dissimilar, the agreement obtained would have been higher. Disagreement regarding Rvar and MI based metrics occurred mainly because selected key terms occurred just once and it was very hard to agree on how such rare terms could be key terms of those documents. We have not achieved to have the results for Czech evaluated by two persons. But it should be mentioned that Czech poses yet another problem when evaluation is at stake, due to its richer morphology. For the example shown in table 1, one observes that multi-words extracted and ranked are mostly sub-phrases of multi-word "group on ethics in science and new technologies" if not of the 11-word term "members of the group on ethics in science and new technologies". While for Portuguese and English this has almost no consequences, for Czech, "skupiny pro etiku ve vědě" is a translation of "of group on ethics in science" which is not exactly a term. Corresponding term in nominative case would be "skupina pro etiku ve vědě". It was accepted as adequate (G) as it also translates as "groups on ethics in science" that is not present in the Portuguese and English versions of the same text. Similarly, "etiku ve vědě" is the accusative case for "etika ve vědě". Results obtained enable however a clear idea about the content of the document. But evaluation, for languages as Czech and other languages having word forms modified by case, still need to be deeply discussed or may require a post extraction normalizer to bring phrases to nominative case.

Table 7: Kappa statistics-based agreement between the evaluators, for Portuguese and English, for Rvar and MI based metrics.

	Portuguese	English
LBM rvar	0.28	0.24
LM rvar	0.27	0.28
LBM MI	0.07	0.28
LM MI	0.19	0.22

6 CONCLUSIONS AND FUTURE WORK

Our approach to key-term extraction problem (of both words and multi-words) is language independent.

By ranking separately words and multi-words, using 20 metrics, based on 4 base metrics, namely Tf-Idf, Phi Square, Rvar (relative variance) and MI (Mutual Information), and by merging top ranked words' list with top ranked multi-words' list taking into account the values assigned to each word and multi-word by each of the metrics experimented we were able to make no discrimination between words and multi-words, as both entities pass the same kind of sieve/metrics to be ranked as adequate key-terms. This way, by comparing 12 metrics, just taking into account word and multi-word based document representation, we could conclude that Tf-Idf and Phi Square based metrics enabled better precision and recall than equivalent precision/recall obtained by Rvar and MI based metrics that tend to extract rare terms. This contradicts results obtained by (Silva & Lopes, 2009, 2010).

As we wanted to extend our methodology to morphologically rich languages, we introduced another document representation in terms of word prefixes and in that way corroborated the conclusions made by (Teixeira, et al., 2011) in their work, where *Bubbled* variants showed interesting results for morphologically rich languages tested.

This other representation led us to the usage of 8 metrics based on the same 4 kernel metrics already mentioned. Experiments were made for Portuguese, English and Czech. Higher precision obtained for Portuguese was obtained using two of the metrics designed to handle prefix, word and multi-word representation. For Czech, and even for English, results were not that spectacular but deserve further attention. As a matter of fact, second best precision for the 5 top ranked key terms candidates, both for Czech and for English was obtained by using Least Bubble Tf-Idf metric.

Again, Tf-Idf and Phi Square derived metrics

were the best performing. Also, it is worth to mention that the Bubble operator enabled some improvements in the results obtained when applied to Rvar and MI metrics. It is worth noticing that, for Portuguese and Czech, for some metrics, precision augmented when we considered top 10 and even top 20 ranked extracted terms in relation to top 5 ranked ones. For Czech that occurred for Least Bubbled Median Tf-Idf and Least Bubbled Median Phi-Square. For Portuguese it was the case for Least Tf-Idf and Least Bubbled Phi-Square.

In future work we will mainly explore Tf-Idf and Phi-square metrics and their derivatives. Then we must take a greater care of the length of texts evaluated. As a matter of fact, for a large text it may make sense an evaluation with 5, 10 or 20 best ranked terms. But for smaller texts taking just the 5 best ranked terms may affect negatively the mean precision of all documents as, in such cases, at most 2 or 3 best ranked terms will probably exhaust good possibilities for document content descriptors.

In what concerns human evaluation we will make an effort for better preparing this work phase in order to overcome evaluation disagreement by discussing the criteria to be used by evaluators while making them explicit.

Regarding the problem identified in section 5 related to having multi-words that are not independent, we must take greater care on this problem, knowing that it is not that easy to solve. Take another example of extracted good descriptors using Phi-Square metric from document 32006D0688 (*in http://eur-lex.europa.eu/en/legis/latest/chap1620.htm*). Below are the terms classified as good:

- asylum
- asylum and immigration
- immigration
- areas of asylum and immigration
- areas of asylum
- national asylum

If we filter out multi-words that are sub multi-words of larger multi-words., in the example above we would have got rid of “asylum and immigration” and “areas of asylum”. But as you see other filtering possibilities might be used. So this must be cautiously addressed. As a matter of fact we are not so sure that a long key term (5-gram) as “areas of asylum and immigration” is a better descriptor than “asylum and immigration”. Equivalently, it might be extrapolated for the example shown in Table 1, that multiword “group on ethics in science and new technologies”, that might be recaptured by binding top ranked multi-words having identical extremities is possibly a good descriptor. But again some care must be taken. If we want to directly extract longer multi-words as that “group on ethics in science and new technologies” we just need to fix the maximum multiword length, this has computational cost. For this work it was fixed at 5.

Concerning Czech, a stricter evaluation would not accept some of the terms that were taken as good as they were case marked and should not be. This will certainly require some language dependent tool filtering. That is more complex than simple lemmatization of words.

In future work, instead of using fixed length character prefixes of words we will pre-process our documents collection to automatically extract real word radicals using some of the existing language independent morphology learners like Linguistica (Goldsmith, 2001) and Morfessor (Creutz & Lagus, 2007).

For Asian languages as Chinese or Japanese, we will apply the extractor (Gomes, 2009) (Silva, et al., 1999) to characters instead of words and extract multi-character, 2-grams and 3-grams, and use single and multi-character strings ranked using the metrics proposed.

For German, the use of language independent morphology learners mentioned above, together with words and multi-words extracted the same way as we did for Portuguese, Czech or English will enable us to extend our methodology to a larger set of languages.

ACKNOWLEDGEMENTS

This was supported by the Portuguese Foundation for Science and Technology (FCT/MCTES) through funded research projects ISTRION (ref. PTDC/EIA-EIA/114521/2009) and VIP-ACCESS (ref. PTDC/PLP/71142/2006).

REFERENCES

Cigarrán, J. M., Peñas, A., Gonzalo, J., & Verdejo, F. (2005). Automatic Selection of Noun Phrases as Document Descriptors in an FCA-Based Information Retreival System. In B. Ganter & R. Godin (Eds.), *ICFCA 2005* (Vol. Lecture Notes in Computer Science 3403, pp. 49-63): Springer Berlin.

Creutz, M., & Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process., 4*(1), 1-34.

Everitt, B. S. (2002). *The Cambridge Dictionary of Statistics* (2 ed.). New York: Cambridge University Press.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistiscs, 27*(2), 153-198.

Gomes, L. (2009). Multi-Word Extractor, from http://hlt.di.fct.unl.pt/luis/multiwords/index.html

Hulth, A. (2003). Improved Automatic Keyword Extraction Given More Linguistic Knowledge *EMNLP '03 Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 216 - 223). Stroudsburg, PA, USA: Association for Computational Linguistics.

Jacquemin, C. (2001). *Spotting and discovering terms through natural language processing*: MIT Press.

Katja, H., Manos, T., Edgar, M., & Maarten de, R. (2009). The impact of document structure on keyphrase extraction *Proceeding of the 18th ACM conference on Information and knowledge management* (pp. 1725-1728). Hong Kong, China: ACM.

Lemnitzer, L., & Monachesi, P. (2008). Extraction and evaluation of keywords from Learning Objects - a multilingual approach *Proceedings of the Language Resources and Evaluation Conference*.

Liu, F., Pennell, D., Liu, F., & Liu, Y. (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 ACL,* (pp. 620–628). Boulder, Colorado: Association for Computational Linguistics.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *An Introduction to Information Retrieval*. Cambridge Cambridge University Press.

Martínez-Fernández, J. L., García-Serrano, A., Martínez, P., & Villena, J. (2004). Automatic Keyword Extraction for News Finder *Adaptive Multimedia Retrieval* (Vol. 3094/2004, pp. 405-427): Springer Berlin / Heidelberg.

Matsuo, Y., & Ishizuka, M. (2004). Keyword Extraction from a single Document using word Co-Occurence Statistical Information. *International Journal on Articial Intelligence Tools, 13*(1), 157-169.

McIlroy, M. D. (2007, Updated April 6, 2010). Suffix arrays, from http://www.cs.dartmouth.edu/~doug/sarray/

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Texts *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 404-411). Barcelona, Spain.

Ngomo, A.-C. N. (2008). Knowledge-Free Discovery of Domain-Specific Multiword Units *Proceedings of the 2008 ACM symposium on Applied computing* (pp. 1561-1565). Fortaleza, Ceara, Brazil: ACM.

Peter, D. T. (2000). Learning Algorithms for Keyphrase Extraction. *Inf. Retr., 2*(4), 303-336. doi: 10.1023/a:1009976227802

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys, 34*(1), 1-47.

Silva, J. F. d., Dias, G., Guilloré, S., & Lopes, J. G. P. (1999). Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. In P. Barahona & J. Alferes (Eds.), *Progress in Artificial Intelligence* (Vol. 1695, pp. 113-132): Springer-Verlag.

Silva, J. F. d., & Lopes, G. P. (1999). A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units *Proceedings of the 6th Meeting on the Mathematics of Language* (pp. 369-381). Orlando.

Silva, J. F. d., & Lopes, G. P. (2009). A Document Descriptor Extractor Based on Relevant Expressions. In S. Lopes, N. Lau, P. Mariano & L. M. Rocha (Eds.), *Progress in Artificial Intelligence* (Vol. 5816, pp. 646-657): Springer-Verlag.

Silva, J. F. d., & Lopes, G. P. (2010). Towards Automatic Building of Document Keywords *COLING 2010 - The 23rd International Conference on Computational Linguistics* (Vol. Poster Volume, pp. 1149–1157). Pequim.

Teixeira, L., Lopes, G. P., & Ribeiro, R. A. (2011). Automatic Extraction of Document Topics. In L. M. Camarinha-Matos (Ed.), *DoCEIS'11 - 2nd Edition of the Doctoral Conference on Computing, Electrical and Industrial Systems* (Vol. 349, pp. 101–108). Caparica, Portugal: IFIP International Federation for Information Processing.

Yamamoto, M., & Church, K. W. (2001). Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus. *Computational Linguistics, 27*(1), 1-30.