Automatic Extraction of Document Topics

Luís Teixeira[1], Gabriel Lopes[2], Rita A. Ribeiro[1]

[1] CA3-Uninova, FCT, Universidade Nova de Lisboa 2829-516 Caparica, Portugal,
[2] DI-FCT/UNL, 2829-516 Caparica, Portugal
[1]{lst,rar}@uninova.pt
[2]{gpl}@fct.unl.pt

Abstract. A keyword or topic for a document is a word or multi-word (sequence of 2 or more words) that summarizes in itself part of that document content. In this paper we compare several statistics-based language independent methodologies to automatically extract keywords. We rank words, multi-words, and word prefixes (with fixed length: 5 characters), by using several similarity measures (some widely known and some newly coined) and evaluate the results obtained as well as the agreement between evaluators. Portuguese, English and Czech were the languages experimented.

Keywords: Document topics, words, multi-words, prefixes, automatic extraction, suffix arrays.

1 Introduction

A topic or a keyword of a document is any word or multi-word (taken as a sequence of two or more words, expressing clear cut concepts) that summarizes by itself part of the content of that document belonging to a collection of documents.

The Extraction of topics (or keywords) is useful in automatic construction of ontologies, document summarization, clustering and classification, and to enable easier and more effective access to relevant information in Information Retrieval. To measure the relevance of a term (word or multi-word) in a document one must take into account the frequency of that term in that document and in the rest of document collection. Desirably, that term should not appear or should be rare in documents focusing on other subject matters.

Tf-Idf, phi-square, mutual information and variance are measures often used to deal with term relevance in documents and document collections ([16] and [1]). In this paper we use those measures (and newly coined variants of them) to extract both single-words and multi-words as key-terms, and compare the results obtained. Additionally, we identify relevant prefixes (with 5 characters length) in order to deal with morphologically rich languages. As no one is able to evaluate prefixes as relevant or non-relevant, we had to project (bubble) prefix relevance into words and multi-words and created, for this purpose, a new operator (bubble) and new relevance measures) to enable the bubbling of prefix relevance, first into corresponding words, and later in multi-words. Simultaneously, we improve discussion started in [1] and continued in [10] and arrive at different conclusions, namely that results obtained by using tf-idf, phi-square and newly derived measures

are better than results obtained by using mutual information or variance and derived measures.

In section 2 we describe how our work contributes to sustainability; related work is summarized in section 3. In section 4 and 5 the measures used are defined; experiments done are described in section 6 and the results obtained are presented in section 7. In section 8 we draw the conclusions on this paper.

2 Contribution to sustainability

This work impacts on sustainability when easy and intelligent access to large document collections is a stake. Our computations use suffix arrays as an adequate data structure and contribute to decrease computing time and power consumption, thus providing new ways to power saving on high performance search centers.

3 Related Work

In [2], [3], and [4] authors propose systems to extract noun phrases and keywords using language depend tools such as stop-words removing, lemmatization, part-of-speech tagging and syntactic pattern recognition. As it will be seen, our work diverges from those ones as it is clearly language independent.

The work in [5] and [6], for multi-word term extraction, rely on predefined linguistic rules and templates to be able to identify certain type of entities in text documents, making them language dependent. In this area, the method proposed in [10] for extracting multi-words, requiring no language knowledge, will be used for extracting multi-words in 3 languages (EN, PT and CZ), as reported in this paper.

In [7] the extraction of Key-words from news data is approached. This is a non-language independent work. A supervised approach for extracting keywords is proposed in [8], using lexical chains built from the WordNet ontology [9], a tool not available for all languages. In [1], the paper that motivated our work, a Key-term extractor (multi-words) is presented together with a metric, the LeastRvar. However, single words are ignored. From the same authors, in [10], the extraction of single and multi-words as key-terms is worked out. However, a share quota for most relevant single and multi-words is predefined, assuming multi-words as better key-terms. In our work, words, multi-words and prefixes are treated identically, with no predefined preferences. Results obtained support this other vision and show that tf-idf and Phi-square-based measures outperform Rvar and Mutual Information based metrics.

4 Measures Used

In this section, for the purpose of completeness, some well-known measures used in this work are presented, as well as those newly coined measures we had to create.

4.1 Known Measures Used

Tf-Idf Metric. Tf-Idf (Term frequency-Inverse document frequency) [1] is a statistical metric often used in information retrieval 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 prefix/word/multiword appears in the document but it is offset by its frequency in the corpus. It should be noticed that we use a probability, $p(W, d_j)$, in equation (1), defined in equation (2), instead of using the usual term frequency factor.

$$\text{Tf-Idf}(W, d_j) = p(W, d_j) * \text{Idf}(W, d_j) . \quad (1)$$

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

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

Where $f(W,d_j)$ denotes the frequency of prefix/word/multiword W in document d_j and Nd_j stands for the number of words of d_j; $||D||$ is the number of documents of the corpus. So, Tf-Idf(W,d_j) will give a measure of the importance of W within the particular document d_j . By the structure of term Idf we can see that it privileges prefixes, multi-words and single words occurring in fewer documents.

Rvar and LeastRvar, two measures based on variance, **w**ere first presented in [1], with the aim of measuring the relevance of multi-words extracted automatically [16], and are formulated as follows:

$$\text{Rvar}(W) = (1 / ||D||) * \Sigma (p(W, d_i) - p(W, .) / p(W, .))^2 . \quad (4)$$

where $p(W, d_j)$ is defined in (2) and $p(W, .)$ is the median probability of word W taking into account all documents. Being MW a multi-word made of word sequence $(W_1 \ldots W_n)$, LeastRvar is determined as the minimum of Rvar() applied to the leftmost and rightmost words of MW.

$$\text{LeastRvar}(MW_i) = \min (\text{Rvar}(W_1), \text{Rvar}(W_2)) . \quad (5)$$

Phi Square Metric. The Phi Square [12] is a variant of the known measure Chi-Square, allowing a normalization of the results obtained with the Chi Square, and is given by the following expression:

$$\varphi^2 = (N \,.\, (AD\text{-}CB)^2 \,/\, (A+C).(B+D).(A+B).(C+D)) \,/\, M \,. \quad \textbf{(6)}$$

Where M is the total number of terms present in the corpus (the sum of terms from the documents that belong to the collection). And where A is the number of times term t occurs in document d; B the number of times that term t occurs in the other documents of the corpus; C stands for the number of terms of the document d subtracted by the amount of times term t occurs in document d; D is the number of times that neither document d or term t occur (i.e. D = N -A -B -C); and N the total number of documents.

Mutual Information. This measure [15] is widely used in language modulation, and its intent is to identify associations between randomly selected terms and in that point determine the dependence that those terms have among them. This measure presented poor results.

4.2 New Measures Used

It was important to have all measures treated the same way. So, if operator "least" was applied to Rvar [1], it should be applied to any other measure used to rank relevance of words, multi-words and prefixes. So, in this section, we describe the newly created measures based on operators "Least" and "Bubbled". In the following, equations consider that MT stands for any of the used measures on this work (Tf-Idf, Rvar, Phi-square or φ^2, and Mutual Information or MI), P a Prefix, W a word, and MW a multi-word made of word sequence ($W_1 \ldots W_n$).

Least Operator. This operator is the same used in the measure LeastRvar, adapted to work with words alone, where we assume that the leftmost and rightmost words of a single word coincide with the word itself.

$$\text{Least\_MT}\ (W) = \text{MT}\ (W) \,. \quad \textbf{(7)}$$

$$\text{Least\_MT}\ (MW) = \text{Min}(\text{MT}(W_1), \text{MT}(W_n)) \,. \quad \textbf{(8)}$$

Bubbled Operator. Another problem we needed to solve was the propagation of the relevance of each Prefix to words having it as a prefix.

$$\text{Bubbled\_MT}\ (W) = \text{MT}\ (P) \,. \quad \textbf{(9)}$$

Having the operators defined we can now present the formulation for the new metrics used.

$$\text{Least\_Bubbled\_MT (W)} = \text{Bubbled\_MT (P)} . \quad (\mathbf{10})$$

$$\text{Least\_Bubbled\_MT (MW)} = \text{Min}(\text{Bubbled\_MT } (W_1), \text{Bubbled\_MT } (W_n)). \quad (\mathbf{11})$$

As in [10] the median of word length in characters was used to better rank words and multi-words, we consider two additional operators: LM for "Least_Median" and LBM, for "Least_Bubbled_Median " defined in (12) and (13), where T represents a term (word or multi-word).

$$\text{LM MT (T)} = \text{Least\_MT (T)} * \text{Median (T)} . \quad (\mathbf{12})$$

$$\text{LBM MT (T)} = \text{Least\_Bubbled\_MT (T)} * \text{Median (T)} . \quad (\mathbf{13})$$

5 Experiments

We worked with a collection of parallel texts, common for the three languages experimented, Portuguese, English and Czech, from European legislation in force (http://eur-lex.europa.eu/). The total number of terms for these collections was of 109449 for Portuguese, 100890 for English and 120787 for Czech.

Multi-words were extracted using LocalMax algorithm [10] as implemented in [13]. SuffixArrays [14] were used for word extraction and for multi-words, words and prefixes counting.

We worked with single words having a minimum length of six characters (this parameter is changeable) and filtered multi-words (with words of any length) removing those containing punctuation marks, numbers and other symbols. Results presented in tables bellow are based on the evaluation of one of the two evaluators, the most critic one. Table 1 shows the top best ranked terms extracted from 3 parallel documents. Tables 2 and 3 show, for the subset of measures used, that were directly evaluated, the average precision obtained for the three languages for one Evaluator.

Evaluators were asked to evaluate 25 best ranked terms for each one of the six measures in those tables. The evaluation assigned a classification (good topic descriptor (G), near good topic descriptor (NG), bad topic descriptor (B), unknown (U), and not evaluated (NE). Last classification (NE) was required because evaluation was indirectly propagated for the rest of measures that were not directly evaluated. K-statistics, used to measure the degree of agreement between evaluators, is shown in table 3, for measures specifically evaluated. Table 4 shows average precision for the N top ranked terms for best evaluated measures with N equal to 5, 10 and 20. In tables 2, 3 and 4, L was used for Least Operator, LM for Least Median Operator, LBM for Least Bubbled Median operator.

6 Results

Some of the results obtained are presented in the following tables.

Table 1. First five terms extracted, ranked accordingly using the measure Phi-Square for all languages for a document in the corpus.

Portuguese	Czech	English
multilinguismo (G)	Mnohojazyčnost (G)	Multilingualism (G)
alto nível sobre o multilinguismo (NG)	Podskupiny (NG)	group on multilingualism (G)
nomeados a título (B)	Mnohojazyčnosti (G)	high level group on multilingualism (NG)
domínio do multilinguismo (G)	Skupiny (NG)	members of the group (B)
composto por oito (B)	vysoké úrovni pro mnohojazyčnost (G)	sub-groups (B)

Table 2. Average Precision for 5 best ranked terms for Evaluator 1 and all Languages.

	φ^2	L tfIfd	LM Rvar	LM MI	LBM φ^2	LBM Rvar
Portuguese	0,723	0,6389	0,463	0,424	0,6222	0,517
English	0,844	0,785	0,472	0,472	0,800	0,524
Czech	0,700	0,750	0,450	0,450	0,550	0,500

Table 3. K-statistics for the two evaluators.

		φ^2	L tfIfd	L M Rvar	LM MI	LBM φ^2	LBM Rvar
K-Statistics	Portuguese	0.552	0.6324	0.11	0.0196	0.635	0.2152
	English	0.7275	0.4375	0.2665	0.2584	0.5786	0.3478

Table 4. Average precision for the best ranked 5,10and 20 terms, for CZ, EN and PT using best applied measures

	Czech			English			Portuguese		
	P(5)	P(10)	P(20)	P(5)	P(10)	P(20)	P(5)	P(10)	P(20)
LM tfidf	0.70	0.65	0.59	0.81	0.78	0.66	0.68	0.63	0.64
LB tfidf	0.80	0.68	0.65	0.85	0.66	0.65	0.86	0.71	0.65
LM φ^2	0.70	0.60	0.58	0.87	0.78	0.70	0.61	0.64	0.59
L φ^2	0.70	0.60	0.58	0.83	0.76	0.69	0.68	0.64	0.59

LBM tfidf	0.65	0.68	0.66	0.82	0.69	0.62	0.83	0.70	0.68
tfidf	0.90	0.86	0.66	0.84	0.74	0.67	0.69	0.70	0.66

7 Discussion

As shown in table 3, agreement between evaluators was higher for the specifically evaluated measures Phi Square, Least Tf-Idf, and Least Bubbled Median Phi-Square. Propagated evaluation to other Tf-idf and φ^2 based measures also showed equivalent agreement results. MI and φ^2 based measures obtained poorer agreement.

Contradicting the point of view presented at [1], we may say that Tf-Idf is a good measure for selecting key-terms. Moreover the terms extracted by both Tf-Idf and Phi-square, or any of its new variants, show better results than the ones obtained by Rvar, or any of its variants, which were considered better than Tf-df in [1].

Rvar and Mutual Information alone were not capable of adequately ranking terms. Only the usage of variants of these measures, applying Least, Bubble and Median operators, improved their results and enabled a selection of best first terms. Otherwise first 200 or 400 terms would be equally ranked.

Evaluated results in table 4 for Portuguese and Czech, two highly inflected languages, are equivalent. Average precision for English is approximately 10% higher than the values obtained for Portuguese or Czech. Best precision results are obtained with different ranking measures for the evaluation of the N best selected key-terms. Tf-Idf alone produces best results for Czech. Least Bubbled Median Tf-Idf is the best for 20 higher ranked key-terms in Portuguese and Czech. Least Median Phi Square and Least Median Tf-Idf works better for English while Least Bubbled Tf-Idf produces better results for Portuguese. Results from variants of Rvar and Mutual Information were always below 55% for all ranges of terms selected (table 2).

Bubbled variants showed rather interesting results for the three languages, especially for Portuguese and Czech. Least and Least Median operators enabled best results for English.

8 Conclusions

Instead of being dependent on specific languages, structured data or domain, we try to approach the key-term extraction problem (of words and multi-words) from a more general and language independent perspective and make no distinctions between words and multi-words, as both kinds of entities pass the same kind of sieve to be ranked as adequate topic descriptors. Also it can be said that the extraction of prefixes (for dealing with highly inflected languages as is Czech and, to a lower degree, Portuguese) and propagating their relevance into words and multi-words, apart from being one of the main innovations presented, enabled high precision (and recall, not shown) values for the top 20 best ranked topic describing terms extracted.

Also the usage of Suffix Arrays has proved to be very efficient and fast in the extraction of words and prefixes from it, also made viable in a more effective way the counting the occurrences of the words, multi-words and Prefixes within the corpus.

9 References

1. J. F. d. Silva, and G. P. Lopes, A Document Descriptor Extractor Based on Relevant Expressions, 14th Portuguese Conference on Artificial Intelligence, EPIA 2009, Aveiro, Portugal, pp. 646-657 (2009).
2. Cigarrán, Joan. M., Anselmo Peas, Julio Gonzalo and Felisa Verdejo, 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. (2005).
3. Liu, Feifan, Deana Pennell, Fei Liu and Yang Liu. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In 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 (2009).
4. Hulth, Anette. Enhancing linguistically oriented automatic keyword extraction. In 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. (2004)

5 Yangarber, Roman and Ralph Grishman. 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 (2000).

6 Jacquemin Christian. Spotting and Discovering Terms through Natural Language Processing. MIT Press, (2001).

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

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

9 Miller, George A. The science of words. Scientific American Library, New York. (1991).

10 J. F. d. Silva, and G. P. Lopes, Towards Automatic Building of Document Keywords, in COLING 2010 - The 23rd International Conference on Computational Linguistics, Pequim, (2010).

11 J. F. d. Silva, G. Dias, S. Guilloré et al., Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units, in 9th Portuguese conference on artificial intelligence Evora, 21-24 September (1999).

12 Everitt B.S., The Cambridge Dictionary of Statistics, CUP, (2002).

13 Multi-Word Extractor, http://hlt.di.fct.unl.pt/luis/multiwords/index.html.

14 Suffix arrays, http://www.cs.dartmouth.edu/~doug/sarray/.

15 Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, An Introduction to Information Retrieval, Cambridge University Press, (2008).

16 Sebastiani, Fabrizio**.** Machine Learning in Automated Text Categorization : ACM Computing Surveys , No. 1, Vols. 34, pp. 1-47, (2002).