Semantic Measures for Keywords Extraction

Davide Colla, Enrico Mensa, and Daniele P. Radicioni(✉)

Dipartimento di Informatica, Università di Torino, Turin, Italy
davide.colla@edu.unito.it, {mensa,radicion}@di.unito.it

Abstract. In this paper we introduce a minimalist hypothesis for keywords extraction: keywords can be extracted from text documents by considering concepts underlying document terms. Furthermore, central concepts are individuated as the concepts that are more related to title concepts. Namely, we propose five metrics, that are diverse in essence, to compute the centrality of concepts in the document body with respect to those in the title. We finally report about an experimentation over a popular data set of human annotated news articles; the results confirm the soundness of our hypothesis.

Keywords: Keywords extraction · Natural language semantics · Conceptual similarity · Word similarity · Lexical resources

1 Introduction

Keywords extraction is a principal task in the analysis of text documents: keywords represent in compact fashion the main topics of document contents, and they are fundamental in a plethora of tasks including information extraction, selection and retrieval. Keywords extraction is a challenging task: it involves analyzing and characterizing documents semantic content—which is a relevant open research problem—, and it has also many applications, in diverse fields such as feature extraction, document filtering and clustering. Furthermore, keywords are customarily used in compiling minimal, dense summaries and in the broader and neighboring field of automatic summarization; they are used to browse document collections, they are beneficial in refining search engine queries and in building contextual advertisements. Attempts to individuate salient textual elements (be them words, phrases or whole sentences) exist that date back to several decades ago [2]. However, despite their relevance and their usefulness for many purposes, explicit keywords are absent from most documents. Providing documents with this meta-information about their content is a costly and time-consuming activity that still requires professionals to manually provide documents with keywords, either chosen from a given thesaurus/taxonomy or based on their own evaluation.

Keyword extraction has been traditionally performed based on lexical information, by adopting support corpora, or controlled vocabularies, moreover, the extraction step has been performed mostly based on statistical methods or on

F. Esposito et al. (Eds.): AI*IA 2017, LNAI 10640, pp. 128–140, 2017.
https://doi.org/10.1007/978-3-319-70169-1_10

machine learning techniques. The analysis is frequently performed at the terms level, having terms relations represented as graphs [11,19].

Our approach differs from those in literature in several aspects: first, we aim at finding salient *concepts* rather than counting *term* frequencies/occurrences. Also, we define concepts relevance as a *relational* feature: our hypothesis is that concepts in a document are relevant in so far as they are semantically connected to the concepts that are present in the title. We define the notion of *centrality*, that can be computed to estimate how tightly the concepts in the body are connected to the title concepts. To compute the conceptual centrality, we propose some novel metrics and some metrics that to the best of our knowledge have never been used before for keywords extraction.

Our approach has the following strengths: it is simple (it basically tests in different manners how conceptual similarity is suitable for keyword extraction); it scales to evolving document collections; and it does not require training the construction of neither *ad hoc corpora* nor controlled vocabularies/thesauri.

The paper is organized as follows: after a brief survey on related work (Sect. 2), we introduce our approach (Sect. 3), by providing full details on the five metrics being proposed (Sect. 3.2). We then report about the experimentation, discuss the obtained results (Sect. 4), and close by elaborating on future work (Sect. 5).

2 Related Work

Several works have been carried out that share some traits with our system. Most approaches to keyword extraction involve three main phases, that are aimed at identifying candidate keywords, at ranking them, and finally at selecting the top ranked ones.

The Rule Discovery System (RDS) uses syntactic information (namely, some given POS patterns such as ADJ-NOUN, NOUN-NOUN, *etc.*) and collects information both internal to the document and collection-based [5]. Ensemble techniques are adopted herein: different classifiers are learned and then combined with a voting mechanism to predict the class associated to a given keyword. In this setting, the keyword extraction task is converted into a binary categorization task, where candidate terms are classified either as keywords or as non-keywords. The RDS is grounded on a pool of features such as term frequency, collection frequency, relative position of the first occurrence, and POS tag. It has been subsequently improved with further filtering of the NP chunks (to eliminate determiners) and with a different choice (more general) of the corpus upon which to compute frequencies.

The authors of TextRank [11] propose an unsupervised graph-based algorithm for both keyword and sentence extraction that leverages a graph representation of the document. Graph nodes contain information on document terms (the algorithm allows for POS-based filtering, and the authors report the best results for nouns and adjectives). The edges are built based on the co-occurrence relation, that accounts for the distance intervening between terms: two vertices are connected if they co-occur within a window of fixed length (2–10 terms). Edges represent an estimation of the cohesion between the terms in the document. In our approach, we use a quite similar measure, the main difference is that we do not collect information about the cohesion between terms but between concepts. Also, the relatedness we measure involves each concept in the document body, and the concepts in the document title.

The co-occurrence of document terms in a graph-based representation is central also in [17], where the relevance of terms is computed on the base of word frequency, word degree, and ratio of degree to frequency. Degree is a measure devised to favor words that occur frequently and in longer candidate keywords. The authors extract one-third keywords w.r.t. the number of words in the graph, as it had been earlier done by [11].

The work [4] proposes the notion of semantic similarity to extract keywords: in this setting, the author exploits a dynamic programming technique for computing Word Sense Disambiguation (WSD). In particular, by referring to the WordNet sense inventory, the maximal similarity between terms is computed, based on the assumption that semantic similarity is an estimator of the strength of the relationship between words. Our work shares some traits with [4]: we also refer to the conceptual level, but we adopt a much broader sense inventory (the vectors in NASARI [1]), and we compute some similarity measures not between each pair of terms in the document, but between terms in the body and in the title.

SemanticRank introduces a method which can be applied to both individual terms and text segments [19]; it relies on the *omiotis* measure, which allows employing the same approach to both keywords extraction and automatic summarization. To compute the semantic relatedness, the SemanticRank algorithm makes use of the WordNet sense inventory paired with a measure based on Wikipedia. The semantic relatedness between two terms is computed by accounting for the path length, for the types of involved edges, and for the depth of the intermediate nodes in the WordNet hierarchy. This measure is then refined through another graph-based formula, that implements one simple and sound intuition: two terms are more related when the number of articles linking to their corresponding Wikipedia pages is higher than the number of articles linking to either of them [19].

3 Semantic Metrics for Keyword Extraction

One major assumption underlying this work is that keywords extraction should be based on the semantic content of documents rather than on the statistics describing terms frequency or terms co-occurrence. As regards as this feature, our approach is close to several of those surveyed above. One difference, in these respects, is in the sense inventory that we use: most previous attempts rely on the WordNet sense inventory, and on similarity measures deriving from those proposed by [15]. We experiment over a much broader sense inventory, namely that of BabelNet [13], and on its vectorial counterpart, NASARI [1].

However, we also hypothesize that only investigating the connections intervening between terms (or concepts) inside the body of a document is not sufficient to individuate its keywords. This is why among the many possible cues that have been proposed in literature, we single out the role of title, and test different measures to investigate in how far the concepts that are expressed in the title may be relevant to extract keywords: to these ends, text documents are represented as a $\langle \mathsf{title}, \mathsf{body} \rangle$ pair.

Furthermore, we focus on documents rather than on documents collections, since they are useful for dealing with collections that change over time, such as news articles directories. Additionally, document-oriented methods "scale to vast collections and can be applied in many contexts to enrich IR systems and analysis tools" [17].

The keywords extraction process has two main phases, the *semantic preprocessing* and the proper *keywords extraction*. The first phase is aimed at individuating the concepts involved in the document, while the latter one is designed to rank them according to some metrics and to select the highest scoring ones.

3.1 Semantic Preprocessing

In this first phase we perform the disambiguation of the document title and body, that is presently carried out through the Babelfy service.[1] The semantic preprocessing allows to filter out stop words (only verbs, nouns and adjectives are retained), permits individuating concepts that are especially frequent in the document being processed (synonyms are rewritten through a single Babel synset ID), and also makes it possible to compare the semantic content conveyed by the title and by the body of the document.

3.2 Keywords Extraction

In the keywords extraction phase, the following steps are performed:

– Matching between body and title concepts, to select the concepts in the body that are most *relevant* to those in the title;
– Keywords are selected as the highest ranked concepts.

Many efforts have been invested to define heuristics to individuate relevant places where typically the more informative terms can be found, for example in the close field of automatic summarization. In this setting, some features have been individuated—since the pioneering work by [2]—as chief factors in conveying document semantic content. Such main features are: *(i)* term frequency, *(ii)* the elements shared between title and body, *(iii)* structural information on the position of such elements within the text, *(iv)* some specific linguistic cues (basically depending on the kind of documents being considered), such as 'In sum', 'For all these factors', *etc.*. However, interestingly enough, it was early found by [2] that term frequency was less relevant than the other mentioned features. Among

[1] http://babelfy.org.

these, we focus on investigating the semantic links between concepts in the title and in the body of documents. In particular, we start from the main assumption

> "that an author conceives the title as circumscribing the subject matter of the document. Also, when the author partitions the body of the document into major sections he summarizes it by choosing appropriate headings. The hypothesis that words of headings are positively relevant was statistically accepted at the 99% of significance" [2, p. 272].

We acknowledge that this assumption only fits documents with title, and does not allow handling some kinds of documents (e.g., novels and narrative in general) where headings may have no title. However, most documents we are interested in (such as scientific articles, news feeds, newspapers articles, goods descriptions, *etc.*) are typically characterized by having titles. Ultimately, we explore simple features obtained by shifting features *(i)* and *(ii)* to a semantic space.

Our control strategy relies on computing the centrality of each concept in the body of documents with respect to the concepts mentioned in its title. The general approach consists in averaging centrality contributions associated to each (body) concept w.r.t. concepts in the title; keywords are then selected by retaining the highest scoring ones.

In detail, the system starts from the lists $T = \{y_1, y_2, \dots, y_L\}$ such that $y \in \mathsf{title}$ and $B = \{x_1, x_2, \dots, x_M\}$ such that $x \in \mathsf{body}$, that contain the Babel synset IDs in the title and in the body, of length L and M, respectively. We then compute the centrality c of the concepts corresponding to the terms x in the body as a function of their semantic relatedness[2] to those in the title:

$$c(x) = \frac{1}{|T|} \sum_{y_i \in T} \text{semrel}(x, y_i). \qquad (1)$$

We devised five metrics that implement the semrel function by exploiting different resources and techniques. Namely, we propose the following metrics: NASARI, NASARIE, UCI, UMASS and $\textsc{ttcs}^{\mathcal{E}}$, that can be arranged into two classes of metrics: those based on NASARI conceptual representations, and those based on coherence measures.

Regardless of the employed metrics, for each document we select as the best keywords those with maximum centrality, that is:

$$Keywords = \underset{x \in B}{\text{argmax}}\ c(x).$$

Using NASARI vectors to compute semantic relatedness. As our first measure, we exploit the semantic vectors of NASARI, that are the vectorial counterpart of BabelNet synsets. Concepts herein (corresponding to a merge of

[2] There is a subtle though neat difference between semantic relatedness and similarity: consider, e.g., that 'eraser' and 'pencil' are related but not similar, whilst 'pencil' and 'pen' are similar.

WordNet synsets with Wikipedia pages) are described through vector representations, whose features are synset IDs themselves. Each such feature is provided with a weight, computed through the metrics of lexical specificity [1]. In the following we will denote the concept identifier by y or x, and the corresponding vector by $\vec{y}$ or $\vec{x}$.

The semantic relatedness between a concept $x \in B$ and the concept $y \in T$ is computed by considering $\rho_x^{\vec{y}}$, that is the *rank* of x in the vector representation for y. More specifically, given two arbitrary elements x and y_i, we compute their relatedness as

$$\text{semrel}(x, y_i) = \left(1 - \frac{\rho_x^{\vec{y_i}}}{length(\vec{y_i})}\right).$$

The rationale underlying this formula is that x is more relevant to concept y_i if x has smaller rank (and heavier weight), that is it is found among the first concepts associated to y_i in $\vec{y_i}$. For example, if we inspect[3] the NASARI vector for the concept door, we find—in decreasing relevance order—that the third term associated to door is window, the tenth wall, the twelfth is lock, and around the hundredth position interior door: the above formula emphasizes the contribution of heavier features, having lower rank.

The centrality of the concept x with respect to each concept $y_i \in T$ can be determined as

$$\text{semrel}(x, y_i) = \begin{cases} 1 & \text{if } \rho_x^{\vec{y_i}} = 1; \\ 0 & \text{if } x \notin \vec{y_i}; \\ \left(1 - \frac{\rho_x^{\vec{y_i}}}{length(\vec{y_i})}\right) & \text{otherwise.} \end{cases}$$

Specifically, in case the concept x is found to have rank 1 for the concept y_i its relevance is supposed to be maximal to the meaning of y_i (it is likely the same term or a close term which is part of the same synset); conversely, in case it is not found in the vector associated to y (thus obtaining $\rho_x^{\vec{y_i}} = 0$), the relatedness (x, y_i) will not contribute anything to the overall centrality of x to the concepts in T.

Using NASARI embed vectors to compute semantic relatedness. We also explored the NASARI Embed version (NASARIE in the following), that contains embedded vector representations of 300 dimensions; the computation of the centrality can be computed in this case by resorting to standard cosine similarity, thus

$$\text{semrel}(x, y_i) = cosSim(\vec{x}, \vec{y_i}).$$

Using UCI coherence measure to compute semantic relatedness. Moreover, we propose two metrics, the UCI measure [14] and the UMass measure [12] that—originally conceived for evaluating Latent Dirichlet Allocation—, have

[3] For the sake of clarity in this example we consider the *lexical* rather than the *unified* vector, i.e. having terms in place of conceptual IDs that are actually used by the system.

been used in the automated semantic evaluation of different latent topic models [18].[4]

Because both the UCI and the UMASS measures natively handle terms rather than concepts, after the semantic preprocessing phase, we need to translate back concepts into terms. However, by exploiting BabelNet, we map all synonyms for a given concept onto a single shared lexicalization, that is chosen as the most common term according to BabelNet counts. This strategy allows reconciling different terms underlying the same sense, thus preserving some semantic trait.

The UCI metrics [14] computes the cohesion between two terms w_1 and w_2 through their pointwise mutual information, that is

$$score(w_1, w_2, \epsilon) = \log \frac{p(w_1, w_2, \epsilon)}{p(w_1)p(w_2)},$$

where the probabilities are estimated by counting word co-occurrence frequencies in a sliding window over an external corpus, such as Wikipedia, Google or MEDLINE,[5] and the ϵ correction is used to ensure that the function always returns real numbers (presently ϵ is set to 1). In our setting, we are interested in computing the cohesion score between the terms in the body and the terms in the title, so that for each concept $x \in B$ lexicalized as w_x and $y_i \in T$ lexicalized as w_{y_i} we compute

$$\text{semrel}(x, y_i) = score(w_x, w_{y_i}, 1).$$

Using UMass coherence measure to compute semantic relatedness. This metrics define a coherence score based on the co-occurrence of the terms w_1 and w_2 as (adapted from [18])

$$score(w_1, w_2, \epsilon) = \log \frac{D(w_1, w_2) + \epsilon}{D(w_2)},$$

where $D(w_1, w_2)$ and $D(w_2)$ count the number of documents containing both w_1 and w_2, and only w_2, respectively. The adopted formula follows the rationale illustrated for the UCI metrics:

$$\text{semrel}(x, y_i) = score(w_x, w_{y_i}, 1),$$

where the concept $x \in B$ is lexicalized as w_x, and $y_i \in T$ is lexicalized as w_{y_i}.

Using the TTCS$^{\mathcal{E}}$ to compute semantic similarity. The last metrics we used in our experimentation relies on a recent lexical resource, the TTCS$^{\mathcal{E}}$, that consists of a vector-based semantic representation. The TTCS$^{\mathcal{E}}$ is compliant with the Conceptual Spaces, a geometric framework for common-sense knowledge representation and reasoning, and contains a novel mixture of common-sense and encyclopedic knowledge [7,10].[6]

[4] In order to compute such measures we used the Palmetto library [16].

[5] Specifically, in the Palmetto implementation, the pointwise mutual information (PMI) and word co-occurrence counts were computed by using Wikipedia as reference corpus [16].

[6] The TTCS$^{\mathcal{E}}$ resource is available for download at the URL http://ttcs.di.unito.it.

Concept representation and similarity computation with the TTCS$^{\mathcal{E}}$. Let D be the set of N dimensions. Such dimensions are relations that report commonsense information like, e.g., IsA, AtLocation, UsedFor, PartOf, MadeOf, HasA, CapableOf, *etc.*. Each concept c_i in the linguistic resource is defined as a vector $\vec{c_i} = [s^i_1, .., s^i_N]$, where each s^i_h constitutes the set of concepts filling a dimension. Each s can contain an arbitrary number of values, or be empty. The TTCS$^{\mathcal{E}}$ can be used to compute the conceptual similarity between concept pairs: specifically, the Symmetrical Tversky's Ratio Model (STRM) has been adopted to compute conceptual similarity [9].

The similarity computed through the TTCS$^{\mathcal{E}}$ is quite different from popular semantic distance measures, that either employ distances between WordNet nodes, or rely on information content measures [15]. One main assumption underlying this approach is that two concepts are similar insofar as they share values on the same dimension, such as when they are both used for the same ends, they share the same components, *etc.*: in this view, e.g., pencil is deemed more similar to pen than to eraser in that both of them are usedFor writing or drawing.

In the present setting, the TTCS$^{\mathcal{E}}$ is employed to compute the conceptual similarity between concepts in the title and those in the documents body according to the formula

$$\text{semrel}(x, y_i) = STRM(\vec{x}, \vec{y_i}),$$

where $\vec{x}$ and $\vec{y_i}$ represent the TTCS$^{\mathcal{E}}$ vectors for the concepts $x \in B$ and $y_i \in T$, respectively.

4 Evaluation

In the last few years several sets of keywords-annotated documents have been collected, annotated and made available, that allow assessing algorithms and their underlying assumptions on scientific articles, news documents, Broadcast News and Tweets (see, for example, [8]).

Dataset. We experimented on the Crowd500 dataset [8], which has been extensively used for testing. The dataset contains overall 500 documents (450 for training and 50 for testing purposes), arranged into 10 classes: Art and Culture, Business, Crime, Fashion, Health, US politics, World politics, Science, Sport, Technology. Documents herein have been annotated by several annotators recruited through the Amazon's Mechanical Turk service. Each keyphrase is provided with a score equal to the number of annotators who selected it as a keyphrase.

Participants. In the following, for the sake of self-containedness, we report the experimental results obtained by [6], where the authors performed a systematic assessment of an array of keyword extractors and online semantic annotators. In particular, we report the results obtained by 2 keyword extractors that participated in the "SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles" (namely, KP-Miner [3] and Maui [20]), and 5 semantic annotators (AlchemyAPI, Zemanta, OpenCalais, TagMe, and TextRazor[7]). With regards to Alchemy, both the keyword extraction (Alch Key) and concept tagging (Alch Con) services were considered. More details can be found in [6].

Experimental setting. We adopted the same setting as in [6], where two experiments have been carried out: in the first one the authors restricted to considering the top 15 keywords for each document in the dataset, while in the second one they considered all annotated keywords. Given the diversity of the metrics employed, some of them typically return a centrality score for each concept in the document (NASARIE, UCI, UMASS), while the other ones (NASARI and TTCS) are only able to express a centrality score for some of the concepts in the document. For this reason, we defined the number of keywords returned by each metrics by considering as minimum the number of keywords having positive centrality score, and as maximum the average of keywords provided for each document in the training set (this figure amounts to 48 keywords per document). Also, since all metrics assessed were used at a conceptual level, our output is mostly composed by individual keywords rather than by keyphrases: accordingly, in the evaluation of the results, we disregarded all keyphrases and focused on the keywords in the gold standard.

Results. The results obtained by testing on the Crowd500 dataset are illustrated in Table 1. Specifically, in Table 1(a) we present the results obtained by comparing the keywords extracted to the top 15 keywords in the Crowd500 dataset, while the results obtained by considering all of the gold standard keywords are provided in Table 1(a). Regarding the first experiment, over the top 15 keywords, we note that in 3 out of 5 of the considered metrics (namely, NASARIE, UCI and UMASS), the F_1 score is higher than those reported in the paper by [6]. Also in the second experiment NASARIE, UCI and UMASS obtained highest F_1 score, whilst the results of NASARI and $\text{TTCS}^{\mathcal{E}}$ are featured by the highest precision.

Discussion. Given the simplicity of the hypothesis being tested (that is: the title-body conceptual coherence is sufficient to individuate the keywords), the adopted metrics performed surprisingly well, and seem to confirm that our hypothesis is sound. We notice that in computing the results over the 15 top ranked keywords (Table 1(a)), the precision of all our measures is quite low, on average half of that obtained by KP-Miner, Maui and TagMe. In any case, this datum would make our metrics inapplicable in a real setting. Although the precision over all keywords (Table 1(b)) is in line with the other systems (except for KP-Miner, that has an advantage of around 10% on our score), the low precision over the first 15 keywords (that are the more relevant ones) shows that the ranking component in the extraction phase must be improved.

[7] Available at the URLs http://www.alchemyapi.com/api/keyword-extraction/, http://developer.zemanta.com/, http://www.opencalais.com/, http://TagMe.di.unipi.it/ and http://www.textrazor.com/, respectively.

Table 1. Results obtained on the test set of the Crowd500 dataset: for each system Precision (P), Recall (R) and F1 Score (F) are reported.

participant	k	P(%)	R(%)	F(%)
Alch Con	15	16.71	2.81	4.82
Alch Key	15	21.63	6.32	9.78
Calais_Soc	15	6.67	0.09	0.17
KP-Miner	15	**41.33**	8.05	13.48
Maui	15	35.87	9.78	15.37
TagMe	15	34.53	11.21	16.93
TxtRaz Top	15	15.78	5.02	7.62
Zem Key	15	29.75	5.15	8.78
NASARI	15	24.89	10.40	14.67
NASARIE	15	15.62	35.47	21.69
UCI	15	16.06	**44.40**	**23.59**
UMASS	15	15.49	42.53	22.71
TTCS$^{\mathcal{E}}$	15	29.08	8.13	12.71

(a) Results on the top 15 keywords in the gold standard.

participant	k	P(%)	R(%)	F(%)
Alch Con	all	16.71	2.81	4.82
Alch Key	all	12.40	16.71	18.24
Calais_Soc	all	13.69	2.60	4.29
KP-Miner	all	40.19	14.46	21.27
Maui	all	27.46	20.30	23.34
TagMe	all	21.02	35.89	26.51
TxtRaz Top	all	6.28	11.52	8.13
Zem Key	all	29.75	5.15	8.78
NASARI	all	39.83	10.86	17.06
NASARIE	all	27.72	36.16	31.38
UCI	all	29.68	**46.28**	**36.17**
UMASS	all	26.76	43.08	33.02
TTCS$^{\mathcal{E}}$	all	**50.36**	8.49	14.54

(b) Results on all keywords.

On the other side, one weakness of our experimentation (which is, admittedly, a preliminary one) is due to the fact that our results do not actually include keyphrases but only keywords, and thus they cannot be directly compared to those of the other systems. We started devising a module for the recognition of Named Entities (which is to date an open problem) to be integrated into the described system. However, even though we were forced to disregard keyphrases, at a closer inspection of the data, in some cases the annotated keyphrases seem to be rather inaccurate: for example, it is frequent to find locutions such as 'video below', 'although people', 'SeaWorld and', 'size allows' and many others.

Finally, by referring to Table 1(b) we note that the traditional trade-off between precision and recall seems to be intertwined with the degree of semantics adopted. In fact, the metrics based on the TTCS$^{\mathcal{E}}$—which is semantically more sophisticated than the other metrics and represents concepts as entities related to other concepts—obtained over 50% precision, whilst the UMASS metrics, which basically counts terms occurrence in documents, obtained 26.76% precision. A full account of the precision over the 10 domains is provided in Table 2: consistently with previous observations and findings, metrics with highest results have higher standard deviation: this fact is trivially explained by the fact that metrics that perform poorly get low scores on most of the domains, which tend to increase their stability [6].

Moreover, in Table 3 we present the number of keywords available on average over the 10 domains, and the actual number of keywords extracted through the considered metrics. These figures have been obtained in the experiment considering all keywords. By comparing the number of keywords returned by TTCS$^{\mathcal{E}}$

Table 2. Analysis of the precision scores by domain (all-keywords experimentation).

Domain	NASARI	NASARIE	UCI	UMASS	TTCS$^{\mathcal{E}}$
Tech	33.92	35.56	31.25	25.00	**60.00**
Sports	**34.05**	18.10	24.99	23.70	28.33
Business	40.29	30.76	27.08	27.50	**50.00**
US politics	38.71	30.63	34.17	32.92	**66.67**
Art and culture	**32.50**	21.95	23.75	22.08	20.00
Science	41.90	26.21	24.58	23.75	**59.58**
Health	33.81	20.39	27.08	22.92	**46.67**
World politics	**68.00**	41.95	46.44	46.44	34.00
Crime	45.12	27.92	27.08	21.25	**60.00**
Fashion	30.04	23.75	30.44	22.08	**78.33**
Median	39.83	27.72	29.68	26.76	50.36
Average	36.38	27.07	27.08	23.73	54.79
STDEV	10.96	7.30	6.73	7.72	18.28

Table 3. Comparison between the average number of keywords actually returned by each metrics, and (first column) the average number of keywords available in the test set.

Domain	DATASET	NASARI	NASARIE	UCI	UMASS	TTCS$^{\mathcal{E}}$
Tech	45	14	43	48	48	2
Sports	26	12	43	45	45	11
Business	37	10	45	48	48	2
US politics	19	5	27	38	38	1
Art and culture	21	5	39	48	48	1
Science	40	20	47	48	48	12
Health	33	14	44	48	48	3
World politics	18	3	20	34	34	9
Crime	37	5	48	48	48	11
Fashion	55	12	48	48	48	11

and NASARI, we observe that even in cases when the TTCS$^{\mathcal{E}}$ returns ‘many’ keywords, its precision still scores high: this is the case, for example, of the domains Sports, Science, Crime and Fashion.

5 Conclusions

In this paper we have explored a novel hypothesis for keyword extraction. It has been designed by starting from the observation that keywords need to be individuated by accessing the conceptual level behind document *lexica*. Building on this tenet, our system performs word sense disambiguation before executing the extraction step. Some of the proposed metrics natively handle concepts (NASARI, NASARIE and TTCS$^{\mathcal{E}}$), while other metrics (UCI and UMASS) require terms, as their statistics are computed at the lexical level.

We have investigated a simple though effective hypothesis: the title provides fundamental (and perhaps sufficient) cues to extract keywords. Different from the literature where basically pools of criteria are investigated at once, we have then proposed five metrics to assess the coherence between documents title and body.

The experimentation showed that our hypotheses are reasonable. Although much work is still needed to improve the quality of the resources we use (in particular for the TTCS$^{\mathcal{E}}$), we obtained results that—as regards as the F_1 score—are competitive with state of the art systems, and show a good performance especially when considering the precision score, which is relevant also for practical uses.

Acknowledgements. We desire to thank Simone Donetti and the Technical Staff of the Computer Science Department of the University of Turin, for their support in the setup and administration of the computer system used in the experimentation.

The authors are also grateful to the anonymous reviewers for their valuable comments and suggestions.

References

1. Camacho-Collados, J., Pilehvar, M.T., Navigli, R.: NASARI: a novel approach to a semantically-aware representation of items. In: Proceedings of NAACL, pp. 567–577 (2015)
2. Edmundson, H.P.: New methods in automatic extracting. J. ACM **16**(2), 264–285 (1969)
3. El-Beltagy, S.R., Rafea, A.: KP-Miner: a keyphrase extraction system for English and Arabic documents. Inf. Syst. **34**(1), 132–144 (2009)
4. Haggag, M.H.: Keyword extraction using semantic analysis. Int. J. Comput. Appl. **61**(1), 1–6 (2013)
5. Hulth, A.: Improved automatic keyword extraction given more linguistic knowledge. In: Proceedings of EMNLP 2003, pp. 216–223 (2003)
6. Jean-Louis, L., Zouaq, A., Gagnon, M., Ensan, F.: An assessment of online semantic annotators for the keyword extraction task. In: Pham, D.-N., Park, S.-B. (eds.) PRICAI 2014. LNCS (LNAI), vol. 8862, pp. 548–560. Springer, Cham (2014). doi:10.1007/978-3-319-13560-1_44
7. Lieto, A., Mensa, E., Radicioni, D.P.: A resource-driven approach for anchoring linguistic resources to conceptual spaces. In: Adorni, G., Cagnoni, S., Gori, M., Maratea, M. (eds.) AI*IA 2016. LNCS (LNAI), vol. 10037, pp. 435–449. Springer, Cham (2016). doi:10.1007/978-3-319-49130-1_32
8. Marujo, L., Gershman, A., Carbonell, J.G., Frederking, R.E., Neto, J.P.: Supervised topical key phrase extraction of news stories using crowdsourcing and coreference normalization. In: Proceedings of LREC, pp. 399–403. ELRA (2012)

9. Mensa, E., Radicioni, D.P., Lieto, A.: MERALI at SemEval-2017 task 2 subtask 1: a cognitively inspired approach. In: Proceedings of SemEval-2017, pp. 236–240. ACL (2017). http://www.aclweb.org/anthology/S17-2038
10. Mensa, E., Radicioni, D.P., Lieto, A.: TTCS$^{\mathcal{E}}$: a vectorial resource for computing conceptual similarity. In: EACL 2017 Workshop on Sense, Concept and Entity Representations and their Applications, pp. 96–101. ACL (2017). http://www.aclweb.org/anthology/W17-1912
11. Mihalcea, R., Tarau, P.: Textrank: Bringing Order into Texts. Association for Computational Linguistics (2004)
12. Mimno, D.M., Wallach, H.M., Talley, E.M., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: EMNLP, pp. 262–272. ACL (2011)
13. Navigli, R., Ponzetto, S.P.: BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. **193**, 217–250 (2012)
14. Newman, D., Noh, Y., Talley, E., Karimi, S., Baldwin, T.: Evaluating topic models for digital libraries. In: Proceedings of the ACM/IEEE JCDL2010. ACM (2010)
15. Pedersen, T., Patwardhan, S., Michelizzi, J.: Wordnet: similarity: measuring the relatedness of concepts. In: HLT-NAACL, pp. 38–41. ACL (2004)
16. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of WSDM 2015, pp. 399–408. ACM (2015)
17. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. In: Text Mining, pp. 1–20 (2010)
18. Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D.: Exploring topic coherence over many models and many topics. In: Proceedings of EMNLP-CoNLL, pp. 952–961. ACL (2012)
19. Tsatsaronis, G., Varlamis, I., Nørvåg, K.: Semanticrank: ranking keywords and sentences using semantic graphs. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 1074–1082. ACL (2010)
20. Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G.: Kea: practical automatic keyphrase extraction. In: Proceedings of JCDL, pp. 254–255. ACM (1999)