Improving LocalMaxs Multiword Expression Statistical Extractor

Joaquim F. Silva[0000−0002−5223−1180] and Jose C. Cunha[0000−0001−6729−8348]

NOVA LINCS, NOVA School of Science and Technology
{jfs,jcc}@fct.unl.pt**

Abstract. LocalMaxs algorithm extracts relevant Multiword Expressions from text *corpora* based on a statistical approach. However, statistical extractors face an increased challenge of obtaining good practical results, compared to linguistic approaches which benefit from language-specific, syntactic and/or semantic, knowledge. First, this paper contributes to an improvement to the LocalMaxs algorithm, based on a more selective evaluation of the cohesion of each Multiword Expressions candidate with respect to its neighbourhood, and a filtering criterion guided by the location of stopwords within each candidate. Secondly, a new language-independent method is presented for the automatic self-identification of stopwords in *corpora*, requiring no external stopwords lists or linguistic tools. The obtained results for LocalMaxs reach Precision values of about 80 % for English, French, German and Portuguese, showing an increase of around $12 - 13$ % compared to the previous LocalMaxs version. The performance of the self-identification of stopwords reaches high Precision for top-ranked stopword candidates.

Keywords: Multiword Expressions · Statistical Extractor · LocalMaxs algorithm · Stopwords.

1 Introduction

Multiword Expressions (MWEs) are sequences of consecutive words in text *corpora*, that is n-grams, having a meaningful content, semantically more or less strong, e.g., the 2-gram "financial crisis", the 3-gram "world population growth", the 5-gram "International Conference on Computer Science". MWEs are useful: *i*) for unsupervised clustering and classification of documents; *ii*) as document keywords; *iii*) for indexing; *iv*) in Statistical Machine Translation. The extraction of MWEs tries to identify semantically relevant n-grams occurring in a *corpus*, by symbolic (morphosyntactic or semantic) or statistical methods [1, 2]. The latter have the advantage of language-independence. The evaluation of MWE relevance is subjective, always made with reference to some context, which may include thematic terms, e.g. "global warming", specific to subject fields, or more general expressions in a language, occurring across multiple domains, still

** This work is supported by NOVA LINCS (UIDB/04516/2020) with the financial support of FCT.IP.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_13

characterised as semantic units. So, the evaluation of the quality of the output of an automatic extractor should be done by a human jury. LocalMaxs multiword statistical extractor [2] is based on two main aspects: *i*) the cohesion between the words within each n-gram; *ii*) a criterion to evaluate the relative cohesion of the n-gram with respect to its neighbourhood. In fact, if the words of an n-gram are cohesive among themselves, then the n-gram is probably semantically strong, and therefore relevant. We propose to improve the Precision of LocalMaxs extractor [2], through a more selective evaluation of the cohesion of each MWE candidate. Apart from that, we present an automatic language-independent stopwords identification method, for general purpose application. In the following we present the improvement to LocalMaxs, the new method for stopwords identification and conclusions. A guide for model reproducibility is in https://github.com/OurName1234/OurFiles/releases/tag/v1.

2 Improvement to LocalMaxs Statistical Extractor

2.1 Background on MWE Statistical Extraction

The foundations of automatic term extraction have been developed for several decades [3–5, 1]. Statistical regularities in natural language texts have been identified, leading to several statistical association/cohesion measures, e.g. MI [5], χ^2 [6], *Dice* [7], SCP [2], *Loglike* [4], c-value, among others, and their application in text processing tasks. Alternatively, language-dependent linguistic or combined linguistic-statistical approaches, e.g. [8], may achieve better quality results. Some methods use Machine Learning approaches. For example, [9] used C4.5 algorithm to classify candidates, but it strongly depends on the high quality of the training data. Concerning the statistical approaches, Xtract [1] identifies collocations reporting around 80 % Precision, using MI measure [5]. In [10], lexical collocations were extracted with *t-score*, reporting about 60 % Precision. Another approach, *mwetoolkit* [11] for MWE extraction provides integration with web search engines and with a machine learning tool for the creation of supervised MWE extraction models if annotated data is available. *Mwetoolkit* uses *t-score*, MI [5], *Dice* [7] and *Log-likelihood* [4] measures, and the highest F1[1] value reported was 30.57 % (56.83 % Precision and 20.91 % Recall).

MWE extraction methods may use stopwords in some step of the extraction pipeline, namely either associated to the *corpora* preprocessing or to the candidate selection. However, unlike the proposal in Sect. 3, most proposals rely on predefined stopwords lists, which may not be available for some languages.

The Cohesion Measures The degree of relevance of an MWE tends to be reflected in the degree of cohesion among its component words. Some widely used cohesion measures, such as $MI(.)$ [5], $\chi^2(.)$ [6], $Dice(.)$ [7] and $SCP(.)$ [2], were originally designed to measure the cohesion between just two consecutive

[1] $F1 = \frac{2.Precision \times Recall}{Precision + Recall}$

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_13

words. The improvement we propose applies to n-grams with more than two words, $(w_1 \dots w_n)$, with $n \geq 2$, following [2] to generalise the cohesion measures.

$$MI\_f((w_1 \dots w_n)) = \log\left(\frac{p(w_1 \dots w_n)}{\frac{1}{n-1}\sum_{i=1}^{i=n-1} p(w_1 \dots w_i)\, p(w_{i+1} \dots w_n)}\right) \quad (1)$$

$$\chi^2\_f((w_1 \dots w_n)) = \frac{\left(N\, f(w_1 \dots w_n) - Avp\right)^2}{Avp\,(N - Avx)\,(N - Avy)} \quad (2)$$

where

$$Avp = \frac{1}{n-1}\sum_{i=1}^{i=n-1} f(w_1 \dots w_i)\, f(w_{i+1} \dots w_n) \quad (3)$$

$$Avx = \frac{1}{n-1}\sum_{i=1}^{i=n-1} f(w_1 \dots w_i) \qquad Avy = \frac{1}{n-1}\sum_{i=2}^{i=n} f(w_i \dots w_n) \quad (4)$$

$$Dice\_f((w_1 \dots w_n)) = \frac{2\, f(w_1 \dots w_n)}{\frac{1}{n-1}\sum_{i=1}^{i=n-1} f(w_1 \dots w_i) + f(w_{i+1} \dots w_n)} \quad (5)$$

$$SCP\_f((w_1 \dots w_n)) = \frac{f(w_1 \dots w_n)^2}{\frac{1}{n-1}\sum_{i=1}^{i=n-1} f(w_1 \dots w_i)\, f(w_{i+1} \dots w_n)} \quad (6)$$

2.2 The Previous Version

LocalMaxs previous version [2] is reviewed in Definition 1.

Definition 1. *Let $W = w_1, \dots w_n$ be an n-gram, $g(.)$ a generic cohesion function and $frq(W)$ the absolute frequency of occurrence of W in a corpus. Let: $\Omega_{n-1}(W)$ be the set of $g(.)$ values of all contiguous $(n-1)$-grams contained in W; $\Omega_{n+1}(W)$ be the set of $g(.)$ values of all contiguous $(n+1)$-grams containing W; $len(W)$ be the length (number of words) of W. W is an MWE if and only if,*

$$frq(W) > 1 \wedge \big(for\ \forall x \in \Omega_{n-1}(W),\ \forall y \in \Omega_{n+1}(W)$$
$$(len(W) = 2 \wedge g(W) > y)\ \vee\ (len(W) > 2 \wedge g(W) > \tfrac{x+y}{2})\big)$$

Thus, previous version of LocalMaxs can be seen as a function $LocalMaxs(g(.))$ parameterised by a cohesion function $g(.)$. The generic function $g(.)$ in Definition 1 can be instantiated with any cohesion measure as long as it is extended for n-grams, with $n \geq 2$. This previous version, besides extracting semantically strong MWE, e.g. "climate change" and "inflation rate in eurozone", it also extracts some n-grams, e.g. "even though", "having established", which, despite frequently co-occurring in *corpora*, are semantically irrelevant. These false MWEs prevent the previous extractor from reaching higher Precision values.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_13

2.3 The Improved Version of LocalMaxs

Two modifications to the criterion for selecting MWEs in LocalMaxs are proposed, as described in Definition 2: *a*) using a generalised mean for the evaluation of the relative cohesion; *b*) using a filtering criterion based on stopwords.

Definition 2. *Let:* $W = w_1, \ldots w_n$ *be an n-gram in a corpus C;* $frq(W)$ *the absolute frequency of occurrence of* W *in* C*. Let function* $LocalMaxs(g(.), p, S)$ *have three parameters:* $g(.)$*, a generic cohesion function; an integer* $p \geq 1$*; and* S*, the set of stopwords in corpus* C*. Let:* $\Omega_{n-1}(W)$ *be the set of* $g(.)$ *values of the two contiguous* $(n-1)$*-grams contained in* W *in* C*;* $\Omega_{n+1}(W)$ *be the set of* $g(.)$ *values of all contiguous* $(n+1)$*-grams that contain* W *in the same corpus;* $len(W)$ *be the length (number of words) of* W*.* W *is an MWE if and only if,*

$$\Big(\big(len(W)=2 \wedge g(W) \geq \max(\Omega_{n+1}(W))\big) \vee \big(len(W)>2 \wedge g(W) \geq \big(\tfrac{\max(\Omega_{n-1}(W))^p+\max(\Omega_{n+1}(W))^p}{2}\big)^{\frac{1}{p}}\big)\Big) \wedge frq(W)>1 \ \wedge w_1 \notin S \wedge \ w_n \notin S$$

The interpretation of the two modifications proposed is as follows.

a) Using a generalised mean: instead of the arithmetic mean in the condition $g(W) > \frac{\max(\Omega_{n-1}(W))+\max(\Omega_{n+1}(W))}{2}$, implicit in Definition 1, we propose the generalized mean in the condition $g(W) \geq \big(\frac{\max(\Omega_{n-1}(W))^p+\max(\Omega_{n+1}(W))^p}{2}\big)^{\frac{1}{p}}$. In this last condition, p is an integer parameter of the extractor. Thus, If $p > 1$, it implies that the cohesion of the n-gram W necessary to consider W as an MWE, tends to be greater than in the case of the arithmetic mean, since for $p > 1$, $g(W)$ will have to be closer to the largest value between $\max(\Omega_{n-1}(W))$ and $\max(\Omega_{n+1}(W))$. In fact, for the arithmetic mean, it suffices that $g(W)$ is superior by an infinitesimal to the value that is at the same distance between $\max(\Omega_{n-1}(W))$ and $\max(\Omega_{n+1}(W))$. Therefore, being more demanding with regard to the value of $g(W)$, this new condition tends to produce fewer False Positives, which is reflected in a higher Precision value of the MWE selected by the extractor. However, this new condition can decrease the Recall value, since some true MWE may not have a sufficient $g(W)$ value to be selected as such.

b) Using a filtering criterion based on stopwords: in a language, usually, a subset of words — called stopwords — is identified, having low semantic content and occurring very frequently in *corpora*, e.g.: "the", "in", "of", in English. This improvement to LocalMaxs considers the judicious location of the stopwords within the MWE candidates to be selected by the extractor. Thus, Definition 2 includes the additional requirement that, for W to be an MWE, its leftmost and rightmost words must not be in the *corpus* stopwords set, that is $w_1 \notin S \wedge w_n \notin S$. This helps rejecting MWE candidates as "even though", "regarded as", "having established" .... In Sect. 2.4, a comparison of results shows the effectiveness of these modifications, LocalMaxs improved version being seen as a function $LocalMaxs(g(.), p, S)$ parameterised by a cohesion function $g(.)$, an integer p representing the exponent of the generalised mean, and the set of stopwords S, used in the selection criterion of the Definition 2.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_13

2.4 Experimental Evaluation of the Improved LocalMaxs Version

The Corpora We used two English *corpora* (EN6.0 Mw and EN0.5 Mw) with 6 019 951 and 500 721 words; and one *corpus* for each of the following languages: French (FR6.1 Mw) with 6 079 056 words, German (DE6.0 Mw) with 6 036 023 words, and Portuguese (PT6.1 Mw) with 6 061 118 words. These were collected from https://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/ .

The Evaluation Criterion The evaluation was made by a jury of three persons. Concerning the Precision of the whole set of n-grams extracted as MWEs by each of the LocalMaxs versions, a large enough random sample, Q, was taken. An n-gram in Q is considered a True Positive MWE, if and only if the majority of evaluators agree. The ratio given by the size of TP (the set of True Positive MWEs) over the size of Q, estimates the Precision. Concerning Recall, a large enough random sample of true MWEs (the *TrREs* set), is obtained by human evaluation from each *corpus*. Let R be the set given by the intersection of *TrREs* with the full set extracted by each LocalMaxs version. The ratio of the size of R over the size of *TrREs* estimates the Recall.

Discussion of Experimental Results Lines under $LocalMaxs(g(.), p, S)$ correspond to Definition 2 with parameters instantiated as described in Table 1 caption. Results show the obtained improvements, where all four cohesion measures present better Precision and Recall than the previous version of the extractor. For all five *corpora*, χ^2_f presents the best values for the combined F1 score metric. The results show that the Precision obtained by the improved algorithm reaches about 80 %, consistently across all considered language *corpora*. Besides the language-independence nature of the algorithm (definitions 1 and 2), the results encourage using this improvement in other languages.

The results show that the replacement of the arithmetic mean with the generalised mean in Definition 2 introduces improvements, as the best result was obtained for $p = 2$ for every cohesion measure. When exponent $p = 1$, the arithmetic mean in Definition 1 is equivalent to the generalised mean in Definition 2. So, the difference between $LocalMaxs(g(.), 1, S)$ and $LocalMaxs(g(.))$ is only due to the restriction that the leftmost and the rightmost word of an MWE can not be a stopword. Then, we conclude that both the aforementioned stopwords restriction and the use of the generalised mean (with $p = 2$) have important contributions to the overall Precision improvement to LocalMaxs, and their orthogonal individual effects add together. In fact, for the example of EN6.0 Mw *corpus* and χ^2_f, we have an increase from 68.5 % (for $LocalMaxs(\chi^2\_f)$) to 74.2 % (for $LocalMaxs(\chi^2\_f), 1, S)$), that is 5.7 % due to stopwords restriction, and another increase from 74.2 % to 80.5 % (for $LocalMaxs(\chi^2\_f, 2, S)$), that is 6.3 % due to $p = 2$. Similar contributions happen to the other language *corpora* as shown in Table 1. Overall, the modifications from Definition 2 lead to a significant improvement in Precision, around 12 − 13 %.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_13

Table 1. Precision and Recall results for each version of LocalMaxs algorithm considering its parameters, for the *corpora* in Sect. 2.4. Lines under the $LocalMaxs(g(.), p, S)$ header present results of the improved version (Definition 2), using four different $g(.)$ cohesion functions, different values for the p exponent, and S instantiated, for each *corpus*, by method in Sect. 3. Results under the $LocalMaxs(g(.))$ header refer to the previous version (Definition 1) for each of the same four cohesion functions.

		$LocalMaxs(g(.), p, S)$									
		EN 6.0Mw		**EN 0.5Mw**		**FR 6.1Mw**		**DE 6.0Mw**		**PT 6.1Mw**	
$g(.)$	p	***Prc***	***Rec***	***Prc***	***Rec***	***Prc***	***Rec***	***Prc***	***Rec***	***Prc***	***Rec***
χ^2_f	**1**	74.2	77.0	73.3	78.5	72.8	75.3	74.3	75.8	73.0	75.0
	2	**80.5**	75.0	**79.3**	75.8	**79.5**	73.8	**80.0**	74.3	**79.8**	73.8
	3	75.0	70.5	74.8	71.0	73.8	69.8	75.0	69.3	75.3	70.0
$Dice\_f$	**1**	63.3	81.5	61.8	82.3	63.5	82.0	62.0	80.3	62.8	82.3
	2	69.3	80.0	68.0	80.8	72.0	81.0	68.8	78.8	69.5	80.3
	3	64.0	76.5	62.8	77.5	65.3	77.3	64.0	78.3	65.0	76.0
MI_f	**1**	69.3	50.5	68.3	52.0	70.3	49.3	69.0	50.8	69.8	50.3
	2	76.8	48.0	75.3	49.5	77.0	47.3	76.0	48.3	75.3	47.0
	3	69.8	44.5	68.3	45.3	69.5	45.5	68.3	46.0	70.3	45.0
SCP_f	**1**	70.0	55.5	70.0	56.8	71.3	54.8	72.0	56.3	71.3	54.0
	2	78.0	55.0	77.5	56.0	78.8	54.0	79.3	53.8	77.8	55.3
	3	71.3	51.0	72.3	51.5	73.0	51.8	71.8	50.8	72.0	50.3
		$LocalMaxs(g(.))$									
		EN 6.0Mw		**EN 0.5Mw**		**FR 6.1Mw**		**DE 6.0Mw**		**PT 6.1Mw**	
		Prc	***Rec***	***Prc***	***Rec***	***Prc***	***Rec***	***Prc***	***Rec***	***Prc***	***Rec***
χ^2_f	–	**68.5**	77.0	**67.3**	78.3	**66.8**	76.8	**67.0**	75.8	**66.5**	76.0
$Dice\_f$	–	60.5	79.5	59.3	80.8	58.8	80.8	58.5	79.8	60.0	79.0
MI_f	–	63.8	50.5	62.5	50.8	64.0	48.3	62.0	48.8	63.3	49.3
SCP_f	–	67.0	55.5	65.8	56.8	65.5	56.0	66.3	55.3	64.8	54.8

3 Identifying the Set of Stopwords in *Corpora*

We propose a new general purpose method for the automatic self-identification of stopwords from each *corpus*. This can be also used to instantiate the set S in Definition 2, preserving its language independence, unlike almost all proposals, which depend on predefined stopwords lists, not always available.

Background on Stopwords Identification To identify the stopwords in a *corpus*, morphosyntactic approaches are language-dependent and pose difficulties in handling multilingual *corpora*, unlike statistical approaches. According to [12], the stopword lists based on Zipf's law are reliable but very expensive to carry out. In [13], to identify stopwords in *corpora*, the authors use the Rocchio classifier and the *IDF* (Inverse Document Frequency) measure, which requires the frequency of each word per document, an information that may not be available. Another proposal [14] identifies stopwords in *corpora* by using Term Frequency (TF). Stopwords are not all equally meaningless and, in [14], are ordered by meaningless according to the used criterion. Using the stopwords lists from

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_13

http://snowball.tartarus.org/algorithms/LANG/stop for several languages, they measure the Precision for several cases of the top n stopwords selected.

A New Stopwords Identification Criterion The proposed stopwords identification criterion combines two factors: the *number of neighbours each word* has in the *corpus*; and the *number of syllables of each word*. Concerning the first factor, number of neighbours of each word, by analysing the occurrence of each word in a *corpus*, it is observed that the higher the number of its distinct left or right neighbouring n-grams, the less meaningful the word is. When considering the number of distinct 2-gram neighbours of each word, this produced better results, by empirical analysis, than any other n-gram size, for the purpose of word meaningfulness. Indeed, large numbers of neighbours are associated with meaningless words; small number of neighbours suggest non stopwords. The second factor, the number of syllables in a word, reflects the effort to pronounce the word. In fact, words with a higher number of syllables tend to occur less often.

The method relies on the combination of the two factors, defining a $NeigSyl(w)$ function, given by dividing the number of 2-gram neighbours each distinct word w has in a *corpus*, by the number of syllables of w. We verified that the higher the value of this quotient, the less meaningful the word. By ordering the words in ranks (r) according to decreasing $NeigSyl(w)$ values, this method allows separating stopwords from content words. Because the list sorted by decreasing $NeigSyl(.)$ values from a *corpus* does not show a clear boundary between stopwords and content words, we need to define a cutoff to automatically separate these two word groups, corresponding to the rank b such that $b = \text{argmax}_r \left(|f(r+\Delta_k) - f(r)| \geq \Delta_k\right)$, where: $f(r) = NeigSyl(word(r))$; $word(r)$ stands for the word corresponding to rank r; Δ_k is a fixed integer distance in the rank ordering. To ignore irregular sharp variations in $NeigSyl(.)$ for neighbouring ranks, we set Δ_k greater than 1; the value leading to the best results was $\Delta_k = 4$.

Experimental Evaluation Using the same reference stopwords lists as in [14], and considering four *corpora*, each one having around 6 Million words (presented in Sect. 2.4), when comparing the Precision values for our criterion *vs* the approach in [14], we obtained the following pairs of values, for the lists of top ranked 50 & 100 stopword candidates: (1.0 & .87) *vs* (0.82 & 0.74) for English; (.96 & .83) *vs* (.74 & .61) for French; (.96 & .82) *vs* (.84 & .74) for German; (.96 & .84) *vs* (.78 & .62) for Portuguese. The proposed criterion achieves a better Precision than using Term Frequency alone [14], showing values around 10 % higher.

4 Conclusions

An improvement is proposed to the LocalMaxs statistical extractor that significantly increases its Precision values, as shown for four natural languages. The proposed method introduces the requirement that an MWE has no stopwords in its leftmost or rightmost words, as well as the replacement of the arithmetic mean

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_13

with the generalised mean in the criterion for evaluating the relative importance of the cohesion of n-grams in the context of their neighbourhoods. Language independence is maintained and the Precision of the LocalMaxs extractor improves by around 12−13 %, reaching about 80 % for the language *corpora* tested. Also, a novel method for automatically identifying the stopwords of each working *corpus* is proposed, which considers both the number of syllables and the number of neighbouring n-grams of each word, leading to improved Precision in the identification of top-n ranked stopwords by language-independent statistical methods, when compared to using Term Frequency only. The method does not depend on predefined stopwords lists, possibly unavailable for some languages. It can also be used as a general tool for monolingual or multilingual *corpora*.

References

1. Frank Smadja. Retrieving collocations from text: Xtract. *Comput. Linguistics*, 19:143–177, 1993.
2. Joaquim Silva, J. Mexia, Carlos Coelho, and Gabriel Lopes. A statistical approach for multilingual document clustering and topic extraction form clusters. *Pliska Studia Mathematica Bulgarica*, 16:207–228, 01 2004.
3. J. Justeson and S. Katz. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering, 1(1)*, 1995.
4. Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.
5. Kenneth Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.
6. William Gale and Kenneth W. Church. in concordance for parallel texts. In *Proceedings of the Seventh Annual Conference of the UW Centre of the new OED and Text Research, Using Corpora*, pages 40–62, Oxford, 1991.
7. Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.
8. Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto. Terminology extraction: An analysis of linguistic and statistical approaches. In *Knowledge Mining*, pages 255–279. Springer Berlin Heidelberg, 2005.
9. Ian Witten, Gordon Paynter, Eibe Frank, Carl Gutwin, and Craig Nevill-Manning. KEA: practical automatic keyphrase extraction. *CoRR*, cs.DL/9902007, 1999.
10. Stefan Evert and Brigitte Krenn. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language*, 19(4):450–466, 2005. Special issue on Multiword Expression.
11. Carlos Ramisch, Aline Villavicencio, and Christian Boitet. mwetoolkit: a framework for multiword expression identification. In *Procs. of the 7th International Conf. on Language Resources and Evaluation (LREC'10)*. ELRA, 2010.
12. Rachel Tsz-Wai Lo, Ben He, and Iadh Ounis. Automatically building a stopword list for an information retrieval system. *J. Digit. Inf. Manag.*, 3:3–8, 2005.
13. Masoud Makrehchi and Mohamed S. Kamel. Automatic extraction of domain-specific stopwords from labeled documents. In *Advances in Information Retrieval*, pages 222–233. Springer Berlin Heidelberg, 2008.
14. Stefano Ferilli, Giovanni Luca Izzi, and Tiziano Franza. Automatic stopwords identification from very small corpora. In *Intelligent Systems in Industrial Applications*, pages 31–46. Springer International Publishing, 2021.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_13