On Frequency-Based Approaches to Learning Stopwords and the Reliability of Existing Resources — A Study on Italian Language

Stefano Ferilli(✉) and Floriana Esposito

University of Bari, Bari, Italy
{stefano.ferilli,floriana.esposito}@uniba.it

Abstract. Natural Language Processing techniques are of utmost importance for the proper management of Digital Libraries. These techniques are based on language-specific linguistic resources, that might be unavailable for many languages. Since manually building them is costly, time-consuming and error-prone, it would be desirable to learn these resources automatically from sample texts, without any prior knowledge about the language under consideration. In this paper we focus on stopwords, i.e., terms that can be ignored in order to understand the topic and content of a document. We propose an experimental study on the frequency behavior of stopwords, aimed at providing useful information for the development of automatic techniques for the compilation of stopword lists from a corpus of documents. The reliability and/or deficiencies of the stopwords obtained from the experiments is evaluated by comparison to existing linguistic resources. While the study is conducted on texts in Italian, we are confident that the same approach and experimental results may apply to other languages as well.

Keywords: Natural Language Processing · Linguistic resources
Stopwords · Keyword extraction

1 Introduction

In spite of the ever-growing spread of multimedia content in digital format, text is still the main channel by which information is represented, exploited and exchanged by humans. Accordingly, the overwhelming majority of content in Digital Libraries (DLs for short) is still in the form of text. In fact, often even non-textual documents are indexed and annotated based on a textual description of their content, also due to the complexity of extracting explicit information from them. It is likely that this landscape will never change significantly, because natural language is the tool that humans have developed and refined through the millenniums to express their thoughts and notions to other people. In turn, the availability of a huge number of texts in DLs and other kinds of digital repositories naturally causes the so-called *information overloading* problem and raises the issue of how to efficiently and effectively manage them, especially

G. Serra and C. Tasso (Eds.): IRCDL 2018, CCIS 806, pp. 69–80, 2018.
https://doi.org/10.1007/978-3-319-73165-0_7

for retrieval and consultation purposes. Indeed, manual management is clearly beyond human capabilities.

In order to solve this problem, research in Computer Science has started several research areas, the most fundamental of which (because all others rely on its results) is Natural Language Processing (NLP), aimed at developing advanced tools for understanding the components, structure and meaning of texts, and to properly organize this information so as to help human users in satisfying their information needs. Typical NLP tasks, ranging from the morphological level, through the lexical one, up to the syntactic and, for some limited applications, the semantic one, are the following:

Language Identification aims at discovering the language in which the text is written;
Stopword Removal removes the terms that are not informative about the specific text content;
Normalization standardizes to a single form inflected forms of words;
Part-of-Speech Tagging associates to each term its grammatical function;
Parsing returns the syntactic structure of sentences;
Understanding aims capturing some kind of semantic information underlying the text.

Each of these steps is typically carried out using suitable linguistic resources. Language Identification often exploits n-gram distribution, Stopword Removal exploits lists of frequent terms, Normalization exploits lists of suffixes (e.g., [18]), Part-of-Speech Tagging exploits suffixes and/or grammatical rules (e.g., [2]), Parsing uses grammars, Word Sense Disambiguation uses conceptual taxonomies or ontologies.

Since languages are very different from each other, these resources are necessarily language-specific. Unfortunately, developing these resources poses several issues. Each language has its own peculiarities, and hence the resources developed for a language are useless for the others. Manually designing and developing these resources by experts is a costly, time-consuming and error-prone activity. Once the resources are available, it is very hard to maintain and update them, or to tailor them to specific domains, or to fix possible errors. Most works in the literature are concerned with English, probably due to its having a structure which is easier than other languages and to its importance as the standard information interchange language worldwide. Little exists for a few other important languages, and almost nothing for the vast majority of minor languages. As a result, automatic high-level processing techniques cannot be applied to documents in these languages, leading to the risk that entire cultures might be lost.

A possible solution is trying to learn the resources and other useful linguistic information (semi-)automatically from texts in the various languages. Some attempts can be found in the literature for Language Identification (in the form of statistics on the distribution of n-grams across the various languages [1,15,16]), Part-of-Speech Tagging (e.g., by learning tagging rules [3,4]), Parsing (with the research stream concerning grammar inference [6]) and Understanding (with initial attempts to learn concept taxonomies or graphs, or even ontologies, but often

based on existing taxonomies/graphs and/or semi-automatically [5,11,13,14,17,21,22]). Our contribution in this landscape was *BLA-BLA* (an acronym for 'Broad-spectrum Language Analysis-Based Learning Application'), a tool that currently includes several techniques that allow to learn in a fully automatic way linguistic resources for language identification [8], stopword removal and term normalization [7,9] and concept extraction [12,19]. The learned resources may be used as returned by the system, and/or be taken as a basis for further manual refinements. Whenever more texts become available for the language, it is easy to run again the technique and obtain updated resources.

Stopwords, in particular, are terms in a language that appear so often and pervasively in the documents as to make them irrelevant to distinguish documents with respect to their content. From an information retrieval perspective, a stopword can be defined as a "word that has the same likelihood of occurring in those documents not relevant to a query as in those documents relevant to the query" [23]. So, by definition, stopwords can be safely ignored by NLP techniques that work at the lexical level. The removal task is simply carried out by look-up in a pre-determined list of words. The usual way for preparing such a list is including all *function words*, i.e. terms associated to invariant Parts-of-Speech of the language (usually articles, pronouns and prepositions), but this requires prior knowledge about the grammar of the language. Moreover, for domain-specific applications, also other terms that are insignificant in the particular context (e.g., the word '*computer*' in a DL specialized in Computer Science) can be added to the list. For instance, [23] adopts this perspective, using a Vector Space Model to identify stopwords. However, the proposed technique applies Porter's stemmer [18] prior to the stopword extraction step, which makes the approach language-dependent, and requires, again, the existence of tools/resources for that language. Two more purely frequency-based approaches are proposed in [10,20]. However, the former still deals with English, and the latter specifically focuses on French. The former was tested on a corpus of broad literature including more than 1 million words, and the latter on two corpora made up of many small texts, but totaling more than 4 and more than 6 million words, respectively. Moreover, both manually adjusts the automatically determined list of stopwords. BLABLA aims at avoiding all these requirements and limitations: it uses just plain texts in a given language for learning, it requires very small corpora, it is fully automatic, it does not focus on a specific language.

In BLA-BLA, stopwords are currently identified as those terms that appear with a higher frequency than the other words in the training documents. The selection is based on a frequency threshold, that in the current prototype is simply set as the average frequency of all terms collected for the language, multiplied by a smoothing factor. So, in this paper, we focus on the automatic learning of stopword lists, with the aim of improving the technique embedded in BLABLA. More specifically, here we present a study of the frequency behavior of terms extracted from texts in a given language, depending on the type and amount of text, and on the mix of texts used for learning. Our study compares the experimental results with standard linguistic resources currently available, and discusses the critical issues

arising from such a comparison, both from the learning perspective and as regards the reliability of the existing resources. The learned lessons are then used to suggest how to possibly improve the approach of BLABLA to learn stopwords, and how to even extract keywords along with stopwords. Section 2 proposes a study of the behavior of frequent words in single texts or small corpora from the perspective of stopword learning. Then, Sect. 3 discusses the outcomes of the study, and Sect. 4 makes a proposal for keyword extraction. Lastly, Sect. 5 concludes the paper and outlines future work directions.

2 Experimental Study

BLA-BLA processes a set of input training documents in pure text, each of which is associated to the corresponding language. It assumes that each document belongs exactly to one language. This does not mean, of course, that it cannot include words or expressions from other languages, but these are to be considered as noise, and suitably handled by the learning approaches.

Concerning the lexical level, a pre-processing step is needed to extract from each document only *words*. In BLABLA, a word is formally defined by the following linear expression pattern:

$$\not\flat P\{W'\}^* W P \not\flat$$

where

- $\not\flat$ is the blank symbol;
- $'$ is the apostrophe;
- $P = \{.|,|;|:|?|!|''|'\}^*$ is a (possibly empty) sequence of punctuation marks;
- $W = \{a|b|\ldots|z\}^+$ is the word (hypothesizing a latin alphabet).

So, a word is a sequence of alphabetic characters only, delimited by blank spaces. Between the initial blank and the first character, and/or between the last character and the final blank, punctuation symbols are allowed (not considered as belonging to the word). The case of an apostrophe joining two words was considered as well.

Once the words only are extracted from a text (or a set of texts) T, they are collected in a multiset W. Then, for each word $w \in W$, its relative frequency is computed as the ratio of its number of occurrences over the overall number of word occurrences in the text(s): $f(w) = k/|W|$, where $|w|_W = k$ (i.e., w has k occurrences in T). Now, members of the *vocabulary* V (i.e., the set of unique words in W) are ranked by decreasing frequency. In a very simple (and simplistic) approach, the problem of determining the stopwords in T may be cast as the problem of cutting this list in such a way that all items above the cut point are considered as stopwords, and all items below the cut point are considered as relevant words. In turn, the cut point may be specified as a frequency threshold $\overline{f}$, such that all words $v \in V$ for which $f(v) \geq \overline{f}$ are considered as stopwords.

For our study, we focused on the Italian language, as an example of a language that has attracted some attention from the NLP community, albeit not as much

as English. So, existing stopword lists for this language may serve as a golden standard on which basing our study. It is also a language having a more complex structure than English, so that one may expect that, if good results are obtained on Italian, then good results might be obtained on most other languages, as well. We wanted to study the case in which only a small corpus is available for learning (say, involving just 10 texts). This should obviously stress the learning approach, because we may expect that, processing a large number of texts, the frequency of real stopwords will in the end become clearly predominant over the other words. We adopted this setting because for some languages (e.g., dialects) only very few written texts are available, because they live mostly in oral conversation.

So, we selected 10 texts from the Project Gutenberg[1] and Liber Liber[2] repositories, which make freely available for download many well-known texts from the literature of several languages. It should be noted that these texts are obtained by applying Optical Character Recognition to scanned images of paper books' pages, and so they contain spelling errors spread through the text, that introduce some noise. This is not necessarily bad, since it allows us to test our approaches also on noisy data, which are what one may expect to have in real-world settings.

Texts were selected so as to ensure a wide range of styles, and to support the study of frequency behavior when increasing the number of texts under different conditions. Specifically, we considered the following texts, where a letter in parentheses identifies the source from which the text was downloaded (G = Project Gutenberg, L = Liber Liber):

La Divina Commedia by Dante Alighieri (G), a poem written in the XIV century;
Codice Civile by the Italian Administration, a technical text of the XX century;
L'Esclusa by Luigi Pirandello (L), a novel written in the second half of the XIX century;
I Promessi Sposi by Alessandro Manzoni (L), a novel written in the first half of the XIX century;
Tutte le novelle by Giovanni Verga (L), a collection of stories written across the XIX and XX centuries;
Passeggiate per l'Italia (volumes 1–5) by Ferdinando Gregorovius (G), a description of travels made in the XIX century.

Table 1 reports some statistics about the length (in number of characters and of words) of the selected texts, and their linguistic variety (column 'Vocabulary' reporting the number of different words in each text). The number of characters and words is approximate (counted by a text editor), while the size of the vocabulary is exact (computed by the pre-processing step).

As the linguistic resource to be used as a golden standard, we chose the stopword list provided by Snowball[3], a well-known tool exploited by many systems

[1] https://www.gutenberg.org/.
[2] https://www.liberliber.it/.
[3] http://snowball.tartarus.org/algorithms/italian/stop.

Table 1. Statistics on the processed texts

#	Text	Chars	Words	Vocabulary
1	La Divina Commedia	561149	97714	12796
2	Codice Civile	1511666	228251	8659
3	L'Esclusa	337589	55846	8919
4	I Promessi Sposi	1307423	220174	19658
5	Tutte le novelle	1591823	264703	21641
6	Passeggiate per l'Italia 1	438868	71467	11995
7	Passeggiate per l'Italia 2	549884	86818	14710
8	Passeggiate per l'Italia 3	478110	75871	12721
9	Passeggiate per l'Italia 4	472272	75618	12183
10	Passeggiate per l'Italia 5	289006	46655	10470
11	Passeggiate per l'Italia	2228140	356429	30855

Table 2. Performance on the processed texts

Text(s) #	1	2	3	4	5	6	7	8	9	10	6–10	All	N-T
P@10	1.0	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
P@20	0.85	0.95	0.95	0.95	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
P@30	0.83	0.87	0.93	0.90	1.0	1.0	0.97	0.93	1.0	1.0	0.97	0.97	0.97
P@40	0.80	0.88	0.85	0.85	0.90	0.95	0.95	0.93	0.93	0.93	0.95	0.95	0.95
P@50	0.74	0.76	0.72	0.80	0.90	0.94	0.92	0.90	0.88	0.92	0.94	0.90	0.90
P@60	0.67	0.68	0.70	0.73	0.85	0.88	0.87	0.90	0.83	0.88	0.90	0.88	0.87
P@70	0.64	0.61	0.69	0.73	0.77	0.83	0.81	0.83	0.81	0.81	0.86	0.83	0.86
P@80	0.60	0.58	0.70	0.70	0.69	0.79	0.76	0.75	0.79	0.76	0.83	0.80	0.81
P@90	0.58	0.54	0.66	0.67	0.66	0.73	0.73	0.73	0.76	0.71	0.78	0.74	0.76
P@100	0.53	0.53	0.62	0.65	0.62	0.73	0.71	0.68	0.71	0.66	0.72	0.72	0.72
P = 1	11	5	14	14	30	33	27	21	33	30	28	27	25
R@100	0.19	0.19	0.22	0.23	0.22	0.26	0.25	0.24	0.25	0.24	0.26	0.26	0.26
P=R@279	0.32	0.28	0.35	0.38	0.38	0.36	0.36	0.36	0.34	0.36	0.37	0.40	0.41

currently in use. In its complete form, it consists of 279 stopwords. Table 2 reports the performance, in terms of precision (P), obtained on each single text (# as per Table 1) and on relevant aggregates of texts ('6–10' = the whole 'Passeggiate per l'Italia'; 'All' = the whole set of texts; 'N-T' = the subset of non-technical texts, excluding 'La Divina Commedia' and 'Codice Civile') when cutting the list at positions that are multiples of 10: so, $P@n$ means precision of the top n items in the ranking (i.e., the percentage of the first n items in the list that are also in the golden standard). These values may be considered as representative of the overall trend, albeit it is worth saying that the behavior of P is not monotonic, so some fluctuations are possible between the reported values. We considered only the first 100 positions both for the sake of readability, and because they represent a 'safety region' where most relevant stopwords should be included.

Table 2 also reports further information. $P = 1$ is the maximum position in the ranking at which 100% precision is preserved. R@100 is the recall value (R) associated to the last position in the 'safety region' (100), to compare it to P@100 and to have an idea of how much of the golden standard is still missing at that point in the list. As a reference, given that the golden standard stopword list includes 279 items, the maximum recall reachable @100 is $100/279 = 0.36$. Finally, $P = R$@279 reports the the performance at position 279. Note that, at this position, precision and recall take the same value ($P = R$).

3 Discussion

Table 2 shows that the length of the text is not strictly related to performance. The best performance is obtained on some volumes of 'Passeggiate per l'Italia', which is written in a kind of journalistic style. A slightly lower, but still quite high performance, is obtained on the stories of 'Tutte le novelle'. The two novels come immediately after, followed by the two texts written using more particular styles, i.e., 'Codice Civile' (technical) and 'La Divina Commedia' (poetry). We may conclude that the writing style counts more than the number of words in the text, which makes sense but was partly unexpected. Specifically, colloquial styles are more useful for finding stopwords than technical ones. As expected, using many texts improves performance. While the improvement may not be outstanding compared to some single texts (e.g., 5 and 6), especially for the upper part of the ranking, a smoother decay in performance is clearly visible, as confirmed by the neat increase in performance @279.

The following is the list of errors, i.e., of words appearing in the top 100 items of each text but not in the golden standard, listed by decreasing frequency:

La Divina Commedia ch sì de d s quel me poi così m là quando già tanto son altro qual occhi ben disse sé lor qui ché or fa né com vidi n ogne elli pur però esser ciò giù altra tal prima ancor poco mondo te sù onde mai;

Codice Civile art può essere seguenti deve diritto cod contratto società caso civ beni disposizioni quando *stato* atto comma cosa parte secondo termine d possono salvo diritti codice legge titolo att devono altri azioni senza norme atti creditore fondo debitore terzo proc ogni valore parti luogo amministratori n persona;

L'Esclusa marta d s così occhi madre *ora* maria quasi no poi me quel sì via due casa signora egli dopo senza anna rocco alvignani ella marito mano *ancora* qua sotto ogni ah prima già disse giorno mani nulla;

I Promessi Sposi d quel s così disse poi renzo cosa de altro due qualche quando *ora* don senza ogni far lucia fatto parte tempo tanto bene gran qui ch altri casa fare dire uomo sempre già dopo;

Tutte le novelle d occhi quel quando senza altro poi *ora* fra due ella s casa tanto *colle colla* sotto ogni disse così cosa mani fatto prima egli capo dopo mano sempre tutta giorno dietro nulla quasi volta *ancora* né;

Passeggiate per l'Italia 1 d città roma *ancora* qui mare castello fra monti s due quando dopo *ora* tempo quasi così perchè campagna poi parte chiesa là strada prima ogni *stato*;

Passeggiate per l'Italia 2 roma d ebrei città chiesa impero tempo s due fra così quando sotto grande *ancora ora* storia tevere ogni parte *stato* già popolo egli quel essa dopo italia papa;
Passeggiate per l'Italia 3 roma d egli città italia così parte tempo *ancora* fra *stato* grande napoleone dopo s ravenna francia due papa essi solo già chiesa avignone *ora* romani quali storia senza quando garibaldi essere;
Passeggiate per l'Italia 4 d napoli città isola s due re mare sicilia quali tutte ogni così dopo fra popolo parte tutta *ancora* capri sotto senza palermo pure grande quasi quando siracusa quel;
Passeggiate per l'Italia 5 così d città s *ora* mare quando arte egli tempo vita perchè sempre già solo *ancora* sicilia intorno ciò due ogni casa tempio cuore allora essa dopo arrio popolo mentre euforione amore verso pompei;
Passeggiate per l'Italia d roma città così s due fra *ancora* tempo egli quando dopo *ora* parte ogni chiesa grande sotto mare quali italia *stato* già qui quel tutte solo senza;
Whole corpus d art s quel quando così può due poi senza altro essere cosa ogni *ora* ch parte tempo dopo prima *stato* occhi disse de tanto altri fatto sì;
Non-technical texts d quel s così quando due poi *ora* senza altro ogni dopo tempo cosa disse *ancora* città tanto egli casa fra prima sempre sotto fatto roma parte.

For the sake of subsequent discussion, we will consider as stopwords all words that do not have a definite meaning, indicating an object or an action, by themselves. According to this perspective, not only articles, pronouns, conjunctions and prepositions are stopwords, but also some adverbs and some verbs (e.g., modal verbs). Based on this definition, we underlined in the previous lists the words that we think are real errors. Words in italics are ambiguous, and can be considered as stopwords or not depending on how one interprets them. For instance, 'stato' may be a noun ('state'), and thus it would not be a stopword, or the past participle of verb 'to be', and thus it would be a stopword. Similarly, 'colla' may mean 'glue' or it may be a contraction of 'con la'; 'colle' may mean 'hill' or it may be a contraction of 'con le'; 'ancora', depending on its accent, may mean 'anchor' or 'still, again'; 'ora' may mean 'hour' or 'now'; etc.

Since the above lists include a very large number of words that we would safely consider as stopwords, a question arises about the reliability and completeness of the state-of-the-art resources currently used for stopword removal (at least, for Italian). Actually, it is quite strange that some of these stopwords are not in the list used as the golden standard. Some examples: 'essere' is the infinitive form of verb 'to be', for which many inflected form are in the list; 'fra' is a very common alternate form of preposition 'tra', which is in the list; etc. More in general, many pronouns and generic adverbs are in these list, but not in the golden standard, even if it does include other similar pronouns or generic adverbs. If the stopwords in the above lists were added to the count of correct items, the results reported in Table 3 would be obtained, where 'Loose' considers the terms in italics as stopwords, and 'strict' considers them as non-stopwords (both the count of such terms, and the resulting precision @100, are reported).

Table 3. Performance on the processed texts: P@100

Text(s)	1	2	3	4	5	6	7	8	9	10	6–10	All	N-T
Original	0.53	0.53	0.62	0.65	0.62	0.73	0.71	0.68	0.71	0.66	0.72	0.72	0.72
Loose	4	30	14	10	7	10	13	15	12	14	8	5	7
P@100	0.96	0.70	0.86	0.90	0.93	0.90	0.87	0.85	0.88	0.86	0.92	0.95	0.93
Strict	0	1	2	1	4	3	3	3	1	2	3	2	2
P@100	0.96	0.69	0.84	0.89	0.89	0.87	0.84	0.82	0.87	0.84	0.89	0.93	0.91

A further insight may reveal other interesting details. From the perspective of texts:

- The poem includes many stopwords in truncated form (e.g., 'd' for 'di', 'ch' for 'che', etc.), which could be expected. Considering as correct these stopwords, the increase in performance would make this text the best one, instead of the worst.
- The technical text includes many specific words, which again could be expected. However, it does not include many stopwords, because they are seldom used in the specific domain. In facts, it still is the worst performing one, even after the corrections are applied.
- Nevertheless, including the technical text in the overall computation yields an improvements over the results obtained on non-technical texts only.
- After corrections, novels become the best-performing non-technical single texts: they reach the same performance as the 'journalistic' text(s) in the strict setting, and are even better than them in the loose setting.

From the perspective of terms/stopwords:

- Wrong terms in the lists associated to sets of texts are pushed towards the end of the list, confirming that larger corpora improve the quality of the results.
- Some terms appear in all lists, suggesting that they are actually stopwords that the golden standard failed to include (e.g., 'd', a truncation of preposition 'di').
- Some terms appear in the majority of lists (e.g., 'quando', 'così' appear in 9 texts; 'dopo', 'due', 'ogni' appear in 8 texts; 'ora', 'ancora' appear in 7 texts; 'già', 'parte', 'quel', 'senza' appear in 6 texts).
- Some terms in italics are in almost all lists, suggesting that they should be considered as stopwords (e.g., 'ora' and 'ancora' appear in 7 texts out of 10).

Finally, interesting considerations may be made also from the perspective of terms that are not stopwords. In facts, it is apparent that they might act as keywords for the corresponding texts: just by reading them one may infer that

- La divina commedia is a poem due to the presence of many truncated words;
- Codice Civile is about regulations and agreements among people;

– 'I Promessi Sposi' and 'L'esclusa' are novels, due to the presence of persons' nouns (their main characters are clearly highlighted, indeed); in particular, L'Esclusa is about family relationships;
– Passeggiate per l'Italia is about geography/landscape, history/politics and art; more precisely, the first three volumes concern Rome, while the last two concern the Reign of the Two Sicilies, including Southern Italy and Sicily.

4 Proposal

Based on the above consideration, our proposal for extending BLABLA by improving its stopword extraction feature and adding a keyword extraction feature is the following. Given a set of texts, the frequency-based approach is used to extract candidate stopwords. If only one text is to be processed, it is likely that the resulting list will contain domain-specific terms, but they might be considered as domain-specific stopwords, according to the literature. If several texts are processed, a comparison of the stopwords extracted from the complete corpus to the stopwords extracted from the single texts may be used both to identify real stopwords and to extract keywords describing the specific content of the single texts. Applying this approach to the above lists would yield the following differences of the words found for the single texts with respect to those found for the whole corpus ('All'):

La Divina Commedia altra ancor ben ché ciò com elli esser fa già già lor là m mai me mondo n né ogne onde or per poco pur qual qui son s s tal te vidi;
Codice Civile amministratori att atti atto azioni beni caso civ cod codice comma contratto creditore debitore deve devono diritti diritto disposizioni fondo legge luogo n norme parti persona possono proc salvo secondo seguenti società termine terzo titolo valore;
L'Esclusa ah alvignani ancora anna casa egli ella giorno già madre mani mano maria marito marta me no nulla qua quasi rocco signora sotto via;
I Promessi Sposi bene casa dire don far fare già gran lucia qualche qui renzo sempre uomo;
Tutte le novelle ancora capo casa colla colle dietro egli ella fra giorno mani mano nulla né quasi sempre sotto tutta volta;
Passeggiate per l'Italia 1 ancora campagna castello chiesa città fra là mare monti perchè quasi qui roma strada;
Passeggiate per l'Italia 2 ancora chiesa città ebrei egli essa fra già grande impero italia papa popolo roma sotto storia tevere;
Passeggiate per l'Italia 3 ancora avignone chiesa città egli essi fra francia garibaldi già grande italia napoleone papa quali ravenna roma romani solo storia;
Passeggiate per l'Italia 4 ancora capri città fra grande isola mare napoli palermo popolo pure quali quasi re sicilia siracusa sotto tutta tutte;

Passeggiate per l'Italia 5 allora amore ancora arrio arte casa città ci cuore egli essa euforione già intorno mare mentre perch pompei popolo sempre sicilia solo tempio verso vita;

Passeggiate per l'Italia ancora chiesa città egli fra già grande italia mare quali qui roma solo sotto tutte.

We think that these results are sensible, especially considering the effort needed to obtain them. Indeed, differently from other techniques for stopword and keyword extraction proposed in the literature, for which heavy computations are required (e.g., to build version spaces based on TF*IDF-like schemes, or to compute statistics about co-occurrences of terms), our approach just requires a simple frequency count and a few set operations on lists of terms.

5 Conclusions and Future Work

Most content in Digital Libraries is still in the form of text, and this predominance will probably never be questioned. Except pure display of these documents, all other tasks are based on some kind of Natural Language Processing, that must be supported by suitable linguistic resources. Since these resources are clearly language-specific, they might be unavailable for several languages, and manually building them is costly, time-consuming and error-prone.

This paper studied the behavior of frequent words in single texts and (small) corpora and, based on the study, proposed a methodology to automatically learn a stopword list for a natural language starting from texts written in that language. The learned list may enable further high-level processing of documents in that language, and/or be taken as a basis for further manual refinements. The study suggested also that relevant keywords may be extracted from the texts with a little extension of the proposed approach. Preliminary experimental results show that the extracted stopwords and keywords are appropriate, and pointed out some deficiencies of standard resources available in the literature.

A future work issue is to define an approach to determine the threshold at which distinguishing stopwords from non-stopwords. Also, a study of the behavior on larger and more varied corpora should be carried out. Finally, an indirect evaluation of the quality of results through the evaluation of the performance of high-level NLP tasks based on the learned resources might be interesting.

References

1. Ahmed, B., Cha, S.-H., Tappert, C.: Language identification from text using n-gram based cumulative frequency addition. Proceedings of Student/Faculty Research Day, CSIS, Pace University, p. 12-1 (2004)
2. Brill, E.: A simple rule-based Part of Speech tagger. In: HLT 1991: Proceedings of the Workshop on Speech and Natural Language, pp. 112–116 (1992)
3. Brill, E.: Some advances in transformation-based Part of Speech tagging. In: Proceedings of the 12th National Conference on Artificial Intelligence (AAAI), vol. 1, pp. 722–727 (1994)

4. Brill, E.: Unsupervised learning of disambiguation rules for Part of Speech tagging. In: Natural Language Processing Using Very Large Corpora Workshop, pp. 1–13. Kluwer (1995)
5. Cimiano, P., Hotho, A., Staab, S.: Learning concept hierarchies from text corpora using formal concept analysis. J. Artif. Int. Res. **24**(1), 305–339 (2005)
6. D'Ulizia, A., Ferri, F., Grifoni, P.: A survey of grammatical inference methods for natural language learning. Artif. Intell. Rev. **36**(1), 1–27 (2012)
7. Ferilli, S., Esposito, F., Grieco, D.: Automatic learning of linguistic resources for stopword removal and stemming from text. Procedia Comput. Sci. **38**, 116–123 (2014)
8. Ferilli, S., Esposito, F., Redavid, D.: Language identification as process prediction using woman. In: Proceedings of the 12th Italian Research Conference on Digital Library Management Systems (IRCDL 2016), p. 12 (2016)
9. Ferilli, S., Grieco, D., Esposito, F.: Automatic learning of linguistic resources for stopword removal and stemming from text. In: Agosti, M., Ferro, N. (eds.) Proceedings of the 10th Italian Research Conference on Digital Library Management Systems (IRCDL 2014), p. 12 (2014)
10. Fox, C.: A stop list for general text. SIGIR Forum **24**(1–2), 19–21 (1989)
11. Hensman, S.: Construction of conceptual graph representation of texts. In: Proceedings of the Student Research Workshop at HLT-NAACL 2004, HLT-SRWS 2004, pp. 49–54. Association for Computational Linguistics (2004)
12. Leuzzi, F., Ferilli, S., Rotella, F.: ConNeKTion: a tool for handling conceptual graphs automatically extracted from text. In: Catarci, T., Ferro, N., Poggi, A. (eds.) IRCDL 2013. CCIS, vol. 385, pp. 93–104. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-54347-0_11
13. Maedche, A., Staab, S.: Mining ontologies from text. In: EKAW, pp. 189–202 (2000)
14. Maedche, A., Staab, S.: The text-to-onto ontology learning environment. In: ICCS-2000 - Eight International Conference on Conceptual Structures, Software Demonstration (2000)
15. Martins, B., Silva, M.J.: Language identification in web pages. In: Proceedings of the 2005 ACM Symposium on Applied Computing, pp. 764–768. ACM (2005)
16. Nagarajan, T., Murthy, H.A.: Language identification using parallel syllable-like unit recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), vol. 1, p. I-401. IEEE (2004)
17. Ogata, N.: A formal ontology discovery from web documents. In: Zhong, N., Yao, Y., Liu, J., Ohsuga, S. (eds.) WI 2001. LNCS (LNAI), vol. 2198, pp. 514–519. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-45490-X_66
18. Porter, M.F.: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980)
19. Rotella, F., Leuzzi, F., Ferilli, S.: Learning and exploiting concept networks with connektion. Appl. Intell. **42**, 87–111 (2015)
20. Savoy, J.: A stemming procedure and stopword list for general french corpora. J. Assoc. Inf. Sci. Technol. **50**, 944–952 (1999)
21. Shamsfard, M., Barforoush, A.A.: Learning ontologies from natural language texts. Int. J. Hum.-Comput. Stud. **60**(1), 17–63 (2004)
22. Velardi, P., Navigli, R., Cucchiarelli, A., Neri, F.: Evaluation of OntoLearn, a methodology for automatic population of domain ontologies. In: Ontology Learning from Text: Methods, Applications and Evaluation. IOS Press (2006)
23. John Wilbur, W., Sirotkin, K.: The automatic identification of stop words. J. Inf. Sci. **18**(1), 45–55 (1992)