2015 Brazilian Conference on Intelligent Systems

The Impact of Contrastive Corpora for Term Relevance Measures

Lucelene Lopes, Paulo Fernandes, Roger Granada and Renata Vieira
PUCRS University – Computer Science Department – Porto Alegre – Brazil
lucelene@gmail.com, paulo.fernandes@pucrs.br, roger.granada@acad.pucrs.br, renata.vieira@pucrs.br

***Abstract*—Estimate the relevance of extracted terms through indices based on contrastive corpora is acknowledged to be efficient. Unfortunately, there is no ground rules to help practitioners and researchers to choose adequate contrastive corpora. In this paper, we present an extensive analysis of different options of contrastive corpora for seven different target corpora. It is our goal to show that the impact of such choice is not as harmful as feared, and such impact tends to be diminished as the number of contrastive corpora increases.**

I. Introduction

The humanity has been producing textual material for more than 5,000 years, and during this time it was never very easy to read everything available about a given subject. The problem during very old times was to find people able to read, but once this problem was massively solved with literacy, the real problem about writing material became how to chose which material was worthy reading.

With the advent of the Internet and the abundance of available data in digital format, the importance of identifying relevant terms within texts became evident. The identification is important not only to index texts, but also to identify concepts of a given domain and eventually generate complex ontological structures. Regardless of the application, to estimate the relevance of terms extracted from corpora is not an easy task. The first problem in estimating term relevance is to define what makes a term relevant to a given corpora.

The oldest and simplest way to estimate term relevance is to observe the term frequency (*tf*), and consequently to assume that the frequency of a term is the only reliable indication of its relevance. Obviously, such simple approach is vulnerable to one single problem: some usual terms may be frequent but not relevant at all. For example, the term "results" is not likely to be relevant to any particular domain, and yet it is very frequent in almost all domains.

To tackle this problem several indices have been proposed. In this paper we are interested in the indices based on contrastive corpora, *i.e.*, indices that take into account the number of occurrences in the target corpus, and also observe the occurrences in other contrastive corpora. A natural difficulty to those indices relies on the choice of contrastive corpora [2]. Therefore, this paper presents an extensive analysis of the impact brought by different sets of contrastive corpora for the *tf-dcf* index [3]. Specifically, we observe all possible combinations of target and contrastive corpora with term lists extracted from seven corpora of different size and domains [1].

Our goal is to illustrate the impact to consider as many contrastive corpora as possible, and also to indicate that the size and specificity of the contrastive corpora is not as relevant as it would be expected. However, it is not the goal of this paper to conduct experiments comparing the results achieved by contrastive corpora based indices with indices solely based on occurrences of the target corpus.

This paper is organized as follows: the next section presents existing measures to estimate term relevance; the third section describes the test bed of our experiments; the fourth section presents and discusses the obtained results; and the Conclusion summarizes our contribution and suggests future work.

II. Existing Measures

The options to estimate the relevance of terms extracted from corpora may be grouped into two large families: the indices that observe only the target corpus, and the indices that observe also contrastive corpora. Single corpus indices can be as simple as term frequency and increase in complexity such as log likelihood [4] and NC-value [5]. Such approaches have been used in Information Retrieval at least since the seventies with the work on term frequency – inverse document frequency (*tf-idf*) [6]. However, new indices following single corpus approach never stopped to appear in the literature, *e.g.*, the domain coherence index published in 2013 [2].

The indices that take into account other (contrastive) corpora are more recent, since these first ideas dated from the late nineties [7], and the first decade of the century have been rich in new indices as the termhood [8]. Among these, the term frequency, disjoint corpora frequency (*tf-dcf*) index has been evaluated as superior to its counterparts [3]. Therefore, in our paper we pay a particular attention to the *tf-dcf* index, described in the following.

A. Chosen Relevance Index - tf-dcf

The basic idea behind *tf-dcf* index is to start from the term frequency in the target corpus (c) and to alter this baseline value taking into account the term frequency in contrastive corpora [9]. Such basic idea is common to all indices using contrastive corpora. The differences brought by the *tf-dcf* formulation resides in:

- to downgrade the baseline value as the term also appears in the contrastive corpora;
- to consider the occurrences in the contrastive corpora in a log scale[1];

[1] Analogously to other indices, any basis of the log scale can be employed to *tf-dcf* index, since regardless of the basis choice the number of occurrences is de-linearized. However, practical *tf-dcf* experiments seem to deliver better results with binary log.

978-1-5090-0016-6/15 $31.00 © 2015 IEEE
DOI 10.1109/BRACIS.2015.16

- to adopt a multiplicative composition of log occurrences in more than one contrastive corpus.

Denoting $\mathit{tf}_t^{(c)}$ the term frequency of term t in corpus c, the *tf-dcf* index for t in the target corpus c considering a set of contrastive corpora $\mathcal{G}$ is formally defined as:

$$\mathit{tf\text{-}dcf}_t^{(c)} = \frac{\mathit{tf}_t^{(c)}}{\prod\limits_{\forall g \in \mathcal{G}} 1 + \log\left(1 + \mathit{tf}_t^{(g)}\right)} \tag{1}$$

This formulation assures that:

- if a term t does not appear in the contrastive corpora ($\forall g \in \mathcal{G} \mid \mathit{tf}_t^{(g)} = 0$), then it has the same value for *tf* and *tf-dcf* ($\mathit{tf\text{-}dcf}_t^{(c)} = \mathit{tf}_t^{(c)}$);
- if two terms have the same term frequency, and the same overall occurrences in contrastive corpora, the term that appears in more contrastive corpora will have a smaller *tf-dcf* index, *e.g.*, a term t_1 with m occurrences in the target corpus c, n occurrences in contrastive corpus g_1, and n occurrences in contrastive corpus g_2 has a smaller *tf-dcf* index than a term t_2 with m occurrences in the target corpus c, $2n$ occurrences in contrastive corpus g_1, and 0 occurrences in contrastive corpus g_2 (see examples in Table I).

TABLE I. EXAMPLE OF *tf-dcf* COMPUTATION CONSIDERING A TARGET CORPORA c AND TWO CONTRASTIVE CORPORA $\mathcal{G} = \{g_1, g_2\}$.

t	$\mathit{tf}_t^{(c)}$	$\mathit{tf}_t^{(g_1)}$	$\mathit{tf}_t^{(g_2)}$	$\mathit{tf\text{-}dcf}_t^{(c)}$		computation
t_1	4	7	0	1.0	=	$4/(4 \times 1)$
t_2	4	1	1	1.0	=	$4/(2 \times 2)$
t_3	4	0	2	2.5	=	$4/(1 \times 1.59)$
t_4	4	0	0	4.0	=	$4/(1 \times 1)$
t_5	8	0	1	4.0	=	$8/(1 \times 2)$
t_6	36	3	3	4.0	=	$36/(3 \times 3)$
t_7	10	0	2	6.3	=	$10/(1 \times 1.59)$
t_8	8	0	0	8.0	=	$8/(1 \times 1)$

In consequence, the computation of *tf-dcf* index for a specific target corpus is naturally impacted by the choice of which contrastive corpora to consider. In fact, all indices based on contrastive corpora are affected by this choice [2]. The basic intuition to be investigated in our current work is that, despite the large impact the choice of contrastive corpora might have to some terms, the overall index behavior tends to be more stable as the number of contrastive corpora increases.

III. TEST BED

To illustrate the impact of different contrastive corpora choices, this section presents the test bed for experiments conducted over seven different corpora and all 448 (7 × 64) possible combinations of target and contrastive corpora[2].

A. The Corpora

Our test bed is composed by seven corpora. These corpora were chosen due to the fact that they are largely used and cover different knowledge areas. It was also important to us to consider corpora of different size and different profiles, being some quite specific and others more generic. A description of each corpus is as follows:

(*a*) Europarl corpus [10] is an extract of the collection of the proceedings of the European Parliament, that comprises in total about 30 million words for each of the 11 official languages of the European Union: Danish, German, Greek, English, Spanish, Finnish, French, Italian, Dutch, Portuguese and Swedish. The bilingual corpus containing sentence aligned texts is freely available at the Statistical Machine Translation site[3].

(*b*) Ohsumed corpus [11] is a test collection created to assist information retrieval research. It is a clinically-oriented MEDLINE subset[4], consisting of 348,566 references (out of a total of over 7 million), covering all references from 270 medical journals over a five-year period (1987-1991).

(*c*) The corpus in the conference organisation domain [12] was constructed to support ontology-related tasks, such as multilingual ontology matching, extension, automatic ontology learning and population in the conference organisation domain. It was built using the multilingual ontology concept labels as seeds for crawling relevant documents from the web through a search engine. In this work, we use the English version of this corpus.

(*d*) TED Talks corpus is a compilation of transcripts from presentations made available by TED, a nonprofit organization [5] that makes available the video recordings together with subtitles in many languages of chosen talks. Almost all talks have been translated by volunteers into about 70 other languages. The collection containing sentence aligned documents is provided by the Web inventory named WIT^3, an acronym for Web Inventory of Transcribed and Translated Talks[13]. The collection contains 1,112 transcribed and translated talks containing topics that span the whole of human knowledge.

(*e*) The Brown Corpus of Standard American English[6] is a general corpus containing about 1 million words of various types of texts, being limited to written American English. It is provided in Natural Language Toolkit (NLTK)[7] and contains 12 subsets of texts, separated into: adventure, belles lettres, editorial, fiction, government, hobbies, humor, learned, lore, mystery, news, religion, reviews, romance and science fiction.

(*f*) The corpus in the Geology domain is composed

[2] The number is 64 is the sum of all possible subsets of the 6 contrastive corpora, *i.e.*, 1+6+15+20+15+6+1.

[3] http://www.statmt.org/

[4] http://ir.ohsu.edu/ohsumed/ohsumed.html

[5] http://www.ted.com

[6] http://icame.uib.no/brown/bcm.html

[7] http://www.nltk.org/

of documents gathered from Geology.com site[8]. The set of documents includes subsets categorized into diamonds, earthquakes, gemstones, general geology, igneous rocks, metamorphic rocks. meteorites, inter alia.

(g) FOOTIE (European Football) corpus [14] was constructed from the transcription of the press conferences scheduled before and after every game played by Italy's national team during the 2008 European football championships (UEFA EURO 2008) held in Switzerland and Austria.

Table II summarizes the characteristics of the seven corpora that were submitted to a term extraction. The name adopted for each corpus (a, b, c, d, e, f, g) was based on a rank from the largest to the smallest.

TABLE II. CORPORA EMPLOYED IN THE EXPERIMENTS

ID	contents	words	terms	specificity
a	Europarl Transcripts	11,624,340	1,864,394	specific
b	Ohsumed	1,803,094	636,766	specific
c	Conference Organization	2,436,168	461,325	specific
d	TED talks Transcripts	553,843	164,899	generic
e	Brown Corpus	255,151	112,352	generic
f	Geology	277,164	93,738	specific
g	European Football	143,255	46,798	specific

The procedure of term extraction was performed in three steps. First each corpus was parsed using the Stanford Lexicalized Parser version 3.3.1 [15]. In this work we used the options "wordsAndTags,penn", where "wordsAndTags" option generates the part-of-speech tagged text, "penn" option generates a context-free phrase structure grammar representation. The second step consists of extracting and refining all noun phrases (NPs) in each corpus. This process starts by reading each syntactic tree generated in the parsed files and extract all terms inside a "NP" structure, since noun phrases are well known for containing conceptual information [16].

After extracting all NPs a refining process is performed where two heuristics are applied to each NP. The first heuristic discards NPs containing symbols other than letters, numbers or hyphen. Terms containing symbols or numbers are probably uninteresting typos or garbage, and likely not meaningful terms to the domain. The second heuristic filters out all determiners (DT) and predetrminers (PDT) from extracted NPs. As determiners express the reference in the context, indicating whether the NP is referring to a definite or indefinite element of a class, to a particular number or quantity, to a closer or more distant element, *etc.*, they do not play an important role in the domain identification. In English, DTs include articles (*e.g.* "the", "a" and "an"), demonstrative pronouns (*e.g.* "this" and "that" and their respective plural forms "these" and "those") and quantifiers (*e.g.* "some", "any", "many"). A PDT is a type of determiner that precedes other DT in a noun phrase. The last step computes the absolute term frequency (*tf*) for each noun phrase in each corpus.

B. Methodology

Once the list of extracted terms and respective number of occurrences in each of the chosen corpora is known, our test methodology consists of three steps:

- for each corpus compute the relevance index (*tf-dcf*) and rank the top 50 terms considering all possible combinations of contrastive corpora, *i.e.*, generate 64 lists of the top 50 ranked terms considering:
  - no contrastive corpora (the relevance index is the term frequency); For example, for corpus a the generated list not using contrastive corpora is denoted:
    $a_{50}[]$
  - one contrastive corpus (it generates 6 lists), for corpus a:
    $a_{50}[b]$ $a_{50}[c]$ $a_{50}[d]$
    $a_{50}[e]$ $a_{50}[f]$ $a_{50}[g]$
  - two contrastive corpora (it generates 15 lists):
    $a_{50}[b,c]$ $a_{50}[b,d]$ $a_{50}[b,e]$
    $a_{50}[b,f]$ $a_{50}[b,g]$ $a_{50}[c,d]$
    $a_{50}[c,e]$ $a_{50}[c,f]$ $a_{50}[c,g]$
    $a_{50}[d,e]$ $a_{50}[d,f]$ $a_{50}[d,g]$
    $a_{50}[e,f]$ $a_{50}[e,g]$ $a_{50}[f,g]$
  - three contrastive corpora (it generates 20 lists):
    $a_{50}[b,c,d]$ $a_{50}[b,c,e]$ $a_{50}[b,c,f]$
    $a_{50}[b,c,g]$ $a_{50}[b,d,e]$ $a_{50}[b,d,f]$
    $a_{50}[b,d,g]$ $a_{50}[b,e,f]$ $a_{50}[b,e,g]$
    $a_{50}[b,f,g]$ $a_{50}[c,d,e]$ $a_{50}[c,d,f]$
    $a_{50}[c,d,g]$ $a_{50}[c,e,f]$ $a_{50}[c,e,g]$
    $a_{50}[c,f,g]$ $a_{50}[d,e,f]$ $a_{50}[d,e,g]$
    $a_{50}[d,f,g]$ $a_{50}[e,f,g]$
  - four contrastive corpora (it generates 15 lists):
    $a_{50}[b,c,d,e]$ $a_{50}[b,c,d,f]$ $a_{50}[b,c,d,g]$
    $a_{50}[b,c,e,f]$ $a_{50}[b,c,e,g]$ $a_{50}[b,c,f,g]$
    $a_{50}[b,d,e,f]$ $a_{50}[b,d,e,g]$ $a_{50}[b,d,f,g]$
    $a_{50}[b,e,f,g]$ $a_{50}[c,d,e,f]$ $a_{50}[c,d,e,g]$
    $a_{50}[c,d,f,g]$ $a_{50}[c,e,f,g]$ $a_{50}[d,e,f,g]$
  - five contrastive corpora (it generates 6 lists):
    $a_{50}[b,c,d,e,f]$ $a_{50}[b,c,d,e,g]$
    $a_{50}[b,c,d,f,g]$ $a_{50}[b,c,e,f,g]$
    $a_{50}[b,d,e,f,g]$ $a_{50}[c,d,e,f,g]$
  - considering all six contrastive corpora, it generates only one list:
    $a_{50}[b,c,d,e,f,g]$
- for each corpus, compute the number of different terms among each pair of the 64 generated lists. It produces a mirrored matrix (element i,j value is equal to element j,i value) with element i,j indicating a number n ($0 \geq n \geq 50$) such that the list represented by row i has $50-n$ equal terms as the list represented by column j, *e.g.*, if element i,j is equal to 4 the list represented by row i has 4 terms different than the list represent by column j. For these matrices the order of lists employed to rows and columns considers increasingly the number of contrasting corpora, and for lists with the same number of contrasting corpora the lexicography order of contrastive corpora is assumed, *i.e.*, for the first corpus (a) the order of sets of contrastive corpora for rows and columns is the order of examples presented before: $a_{50}[], a_{50}[b], \ldots, a_{50}[c,d,e,f,g], a_{50}[b,c,d,e,f,g]$.

[8] http://geology.com

- for each corpus compute the average improvement statistically significant achieved by the increase of one more contrastive corpus, *i.e.*, the statistical significance of the difference between pair of lists for the number of terms from 0 to 1 contrastive corpora, from 1 to 2, from 2 to 3, from 3 to 4, from 4 to 5, and finally from 5 to 6.

IV. Results

In this section we present and analyze the produced results with two different points of view. A qualitative analysis of the terms cast out of the top 50 list by using contrastive corpora, and a qualitative analysis of the changes brought increasing the number of contrasting corpora.

A. Term Qualitative Analysis

The first analysis is made observing the lists of Geology domain (corpus f). Not using contrastive corpora the top 50 terms are listed in Table III first column ($f_{50}[]$), while the second column lists the top 50 terms ranked using the other 6 corpora as contrastive ($f_{50}[a, b, c, d, e, g]$).

Both lists show some problems associated to the extraction procedure, namely, the lack of knowledge to recognize as same terms, those written with upper and lowercase letters, but also singular and plural versions. Examples of such problems are the terms “natural gas” and “Natural Gas” which are not considered equal, nor the terms “area” and “areas”.

In Table III the terms “Interstate” and “USGS” are marked in bold since they are the only two that were not cast out of the top 50 ranked ones by using the 6 contrastive corpora. Such large number (48 term changes) shows an undeniable effect brought by *tf-dcf* index. It is, however, necessary to form an opinion about the positiveness of such effect. In order to have a unbiased opinion on that, we conducted a classification of the extracted terms consulting three Geologists to define whether a term was relevant to Geology.

Observing the 48 terms present only in the left hand side, we notice quite generic terms, and one fourth (12 terms) that, even though being general, are naturally related to Geology (terms marked with $\star$). Therefore, three fourths (the 36 unmarked terms) are clearly unspecific to the Geology domain, *i.e.*, casting this terms out was a good effect brought by the use of *tf-dcf* index.

On the contrary, observing the 48 terms appearing only at the right hand side, just 4 terms (one twelfth) were not specific to Geology domain (terms marked with •). Such clear predominance (44 out of 48) of specific terms while using 6 contrastive corpora indicates the quality improvement brought by *tf-dcf* index and a large number of contrastive corpora.

In summary, we will accept as fact that using *tf-dcf* and 6 contrastive corpora improves the quality of extracted terms. More than that, we will assume that every time there is a change in the top 50 term lists due to adding a contrastive corpora, such change is beneficial. Therefore, comparing a list of top 50 terms generated using x contrastive corpora with another list generated using y contrastive corpora, assuming $x > y$, it will be assumed that the better quality of the first list over the second one will be proportional to the number of different terms between the two lists. For example, the list with the top 50 terms for the Geology corpus (corpus f) using contrastive corpora a and b, denoted $f_{50}[a, b]$, and the list obtained with contrastive corpora c, d and e ($f_{50}[c, d, e]$) have 22 different terms between them, and we will assume that $f_{50}[c, d, e]$ is 44% (22/50) better than $f_{50}[a, b]$.

TABLE III. Top 50 terms for Geology corpus

$f_{50}[]$	$f_{50}[a, b, c, d, e, g]$
area	Aerials Online Topo Mapping
areas	agate
article	basalt
Australia	• Beginners Guide
beds	county boundaries
coal	Crater
com	debris flows
country	Detailed topographic maps
$\star$ Earth	Eagle Ford Shale
$\star$ eruption	East-West interstates
feet	elevation trends
$\star$ gas	frac sand
$\star$ Geology	gemstone
$\star$ geology	gemstones
gold	generalized topographic map
image	gneiss
inches	igneous
information	**Interstate**
Interstate	lode
lakes	major physical features
land	major physical features of state
map	map detail
miles	Map Dimensions
$\star$ minerals	Marcellus Shale
NASA	Mauna Kea
$\star$ natural gas	NASA Earth Observatory
number	Natural Gas
$\star$ oil	natural gas liquids
others	natural gas-to-liquids plant on Gulf coast
part	• nice views
parts	• nice views of state
people	North-South interstates
point	opal
production	organic remains
$\star$ rock	quartz
$\star$ rocks	Raven Maps
Route	Satellite Image
$\star$ sand	Satellite Image Map
size	Sea Level
species	south routes
state	stream gages
$\star$ surface	stream levels
time	tsunami waves
United States	United States Geological Survey
use	**USGS**
USGS	Utica Shale
water	Viewing Landsat Images
world	Wall Map Custom Printed Topos Custom
year	west routes
years	• Zoom

B. Systematic Quantitative Analysis

To illustrate the effect of using different combinations of contrastive corpora, Fig. 1 presents graphically the number of different terms found between pairs of top 50 term lists. In this figure, there is one matrix to each corpora (from a to g), plus one matrix with the average number of different terms considering the seven corpora matrices. In each matrix the rows and columns indicate a possible use of contrastive corpora for a target corpus. For instance, in the matrix referring to corpus a the element in row “b c d” and column “d e f g” indicates the number 18, meaning that list $a_{50}[b, c, d]$ has 18 terms that are not present in list $a_{50}[d, e, f, g]$ (and vice versa).

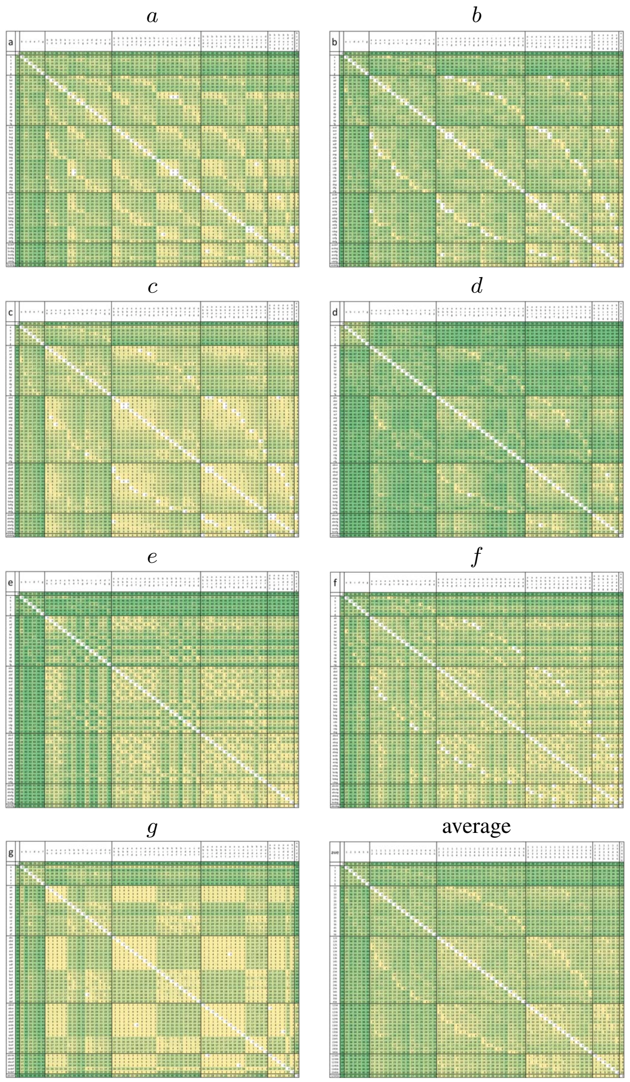

Fig. 1. Matrices with the number of different terms between pairs of lists.

For each target corpus the matrix was build ordering its 64 possibilities of contrastive corpora using lexicographic order and number of contrastive corpora. Obviously, it is not expected to the readers to read in detail all these numeric information, since it is depicted in this figure 32,768 values, *i.e.*, 64^2 numbers in each of the seven matrices, plus one matrix with the average number obtained for all matrices. Hence, we indicate with colors each matrix element, going chromatically from a dark green for higher values (starting with 50) to a light yellow for lower values (ending with 0 - which is found in the diagonals). Such colorful representation let us observe the matrices in Fig. 1 as patterns of the effect brought by using different options of contrastive corpora.

Observing the patterns in Fig. 1, the common point among all matrices is the fact that the higher values are encountered towards the left upper corner, while smaller values appear towards the right lower corner. Such behavior is more pronounced in some matrices as those referring to corpora d and e, but it is present in all matrices. This characteristic indicates that increasing the number of contrastive corpora keep changing the terms in the top 50 lists. However, this change becomes less pronounced as the number of contrastive corpora increases. In other words, using one contrastive corpus bring many changes, and each new added contrastive corpus has a smaller impact. It is interesting to observe that this was verified regardless of the size of the corpora, or its specificity.

Another interesting observation from the patterns shown in Fig. 1 is that there is a group behavior considering the same number of contrastive corpora. In these matrices the quadrants referring to the same number of contrastive corpora are marked by bold frames. Within each quadrant, we observe particular patterns consistent with specificities of each corpora. Such phenomenon is clear in matrices for corpora a and g, but for all corpora similar patterns were found.

Pushing the analysis further, a study on the statistical significance of increasing contrastive corpora was made to each target corpora. Specifically, we consider all number differences between pairs of lists obtained using n and $n+1$ contrastive corpora as samples. Using Spearman rank-order correlation [17], we obtain the values for each target corpus, as denoted in Table IV. In this table is indicated one column to each target corpus, with a last column considering the overall value considering all target corpora as samples. The rows indicate the average gain, in number of terms, brought by each contrastive corpus addition.

TABLE IV. Term change brought by adding contrastive corpora.

corpora	target corpus							overall
from	a	b	c	d	e	f	g	values
0 to 1	26.33	34.50	26.50	22.33	37.17	34.83	29.50	30.17
1 to 2	22.42	30.51	20.98	36.57	39.28	27.41	26.37	29.08
2 to 3	16.77	23.96	15.67	34.78	24.17	19.47	15.13	21.42
3 to 4	14.67	20.30	11.35	27.83	16.10	14.40	11.39	16.57
4 to 5	10.54	14.17	8.30	18.11	10.58	9.82	8.18	11.39
5 to 6	5.33	5.33	4.33	8.67	5.17	4.67	8.67	6.02

Fig. 2 graphically depicts the results for all seven target corpora. Observing this figure it is noticeable that adding a contrastive corpora is always beneficial. However, as the number of contrastive corpora increases the benefits tends to be smaller. Specifically, for corpora a, b, c, f and g this behavior was always consistent, *i.e.*, the higher benefit is brought using one contrastive corpus, adding a second contrastive corpus improves the result, but the benefit brought is less significative, and so on until using six contrastive corpora.

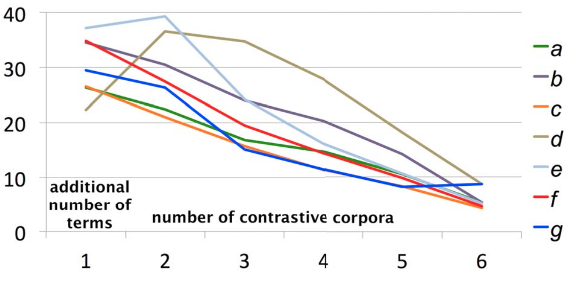

Fig. 2. Term change brought by adding contrastive corpora.

For target corpus d the behavior is different since the addition of a second contrastive corpus brings more benefit than the use of the first one. This same behavior was also found for target corpus e, even though for d it is more pronounced.

It is noticeable that these two corpora (d and e) are not domain corpora, since d has transcripts of TED talks, which cover several domains, and e is the whole Brown corpus also covering several domains. Despite that, the different behavior found in these two corpora was noticeable only comparing the improvement brought by the second contrastive corpora which was greater than the improvement brought by the first corpora. For the benefits brought by the third, fourth, fifth and sixth contrastive corpora all seven target corpora have shown a similar behavior.

It is important to stress that the values indicated are statistically significant improvements brought by adding a contrastive corpus. Therefore, adding contrastive corpora is always beneficial, and a smaller value still indicates an improvement.

V. Conclusion

The use of contrastive corpora to better estimate the relevance of extracted terms is generally accepted, but some authors intuitively argue that such technique is difficult to be employed since there is no general rules to choose the right contrastive corpora set [18], [3], [2]. Our present study offers a counter argument in terms that using contrastive corpora is always beneficial, and piling up more contrastive corpora may not bring as much benefits as the first contrastive corpora, but it will never be harmful. More than that, we did not observe different behavior for corpora with different sizes, or with clear focus or not. It is important to remember that our experiments were conducted over clearly defined domain corpora, as Geology (f), but also quite generic ones as TED talks (d) and Brown corpus (e).

As mentioned, we conducted an analysis of the extracted terms quality only for the Geology corpus, which is probably the clearer focused corpus in our collection. To do so, we consult three Geologists in order to define whether a term was generic or specific to Geology. It was noticeable that even the expert opinions were far from unanimous, and consequently we believe that a more formal analysis of terms could bring a little more confidence in the quality of term lists. However, the focus of our analysis was not the good quality of the adopted index (*tf-dcf*), which was assumed, but the effect brought by choosing different options of contrastive corpora.

In such way, we claim that the choice of contrastive corpora, even though being relevant for fine tuning of extracted term lists, is not a matter of argument to discard the use of contrastive corpora based indices. As we observed, the use of contrastive corpora, as many as possible, will always represent an advantage. Such conclusion does not prevent us to suggest further study on the matter of contrastive corpora, as for instance, to analyze the impact of contrastive corpora choices with respect to practical applications of term ranking. Our goal is to present some evidences to the research community and to encourage the adoption of contrastive corpora-based technique and indices without a fear to badly choosing contrastive corpora, since the worst choice would not use it.

References

[1] L. Lopes and R. Vieira, "Heuristics to improve ontology term extraction," in *PROPOR 2012 – International Conference on Computational Processing of Portuguese Language*, ser. LNCS vol. 7243, 2012, pp. 85–92.

[2] G. Bordea, P. Buitelaar, and T. Polajnar, "Domain-independent term extraction through domain modelling," in *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence*, ser. TIA 2013. Paris, France: Université Paris Nord, 2013, pp. 61–68. [Online]. Available: http://www.insight-centre.org/sites/default/files/publications/tia2013.pdf

[3] L. Lopes, P. Fernandes, and R. Vieira, "Domain term relevance through tf-dcf," in *Proceedings of the 2012 International Conference on Artificial Intelligence (ICAI 2012).* Las Vegas, USA: CSREA Press, 2012, pp. 1001–1007.

[4] A. Gelbukh, G. Sidorov, E. Lavin-Villa, and L. Chanona-Hernandez, "Automatic term extraction using log-likelihood based comparison with general reference corpus," in *Natural Language Processing and Information Systems*, ser. Lecture Notes in Computer Science, C. Hopfe, Y. Rezgui, E. Mtais, A. Preece, and H. Li, Eds. Springer Berlin Heidelberg, 2010, vol. 6177, pp. 248–255. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-13881-2_26

[5] K. Frantzi, S. Ananiadou, and J. Tsujii, "The c-value/nc-value method of automatic recognition for multi-word terms," in *Research and Advanced Technology for Digital Libraries*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 1998, vol. 1513, pp. 585–604. [Online]. Available: http://dx.doi.org/10.1007/3-540-49653-X_35

[6] K. Spärck-Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[7] K. Kageura, "Theories "of" terminology: a quest for a framework for a study of term formation," *Terminology*, vol. 5, no. 1, pp. 21–40, 1999.

[8] C. Kit and X. Liu, "Measuring mono-word termhood by rank difference via corpus comparison," *Terminology*, vol. 14, no. 2, pp. 204–229, 2008.

[9] L. Lopes, "Extração automática de conceitos a partir de textos em língua portuguesa," Ph.D. dissertation, PUCRS University - Computer Science Department, Porto Alegre, Brazil, 2012.

[10] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT summit*, vol. 5, 2005.

[11] W. Hersh, C. Buckley, T. Leone, and D. Hickam, "Ohsumed: An interactive retrieval evaluation and new large test collection for research," in *SIGIR94*. Springer, 1994, pp. 192–201.

[12] R. Granada, L. Lopes, C. Ramisch, C. Trojahn, R. Vieira, and A. Villavicencio, "A comparable corpus based on aligned multilingual ontologies," in *Proceedings of the First Workshop on Multilingual Modeling*, ser. MM '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 25–31. [Online]. Available: http://dl.acm.org/citation.cfm?id=2392696.2392700

[13] M. Cettolo, C. Girardi, and M. Federico, "Wit^3: Web inventory of transcribed and translated talks," in *Proceedings of the 16^{th} Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[14] A. Sandrelli, "Introducing footie (football in europe): simultaneous interpreting in football press conferences," in *Breaking Ground in Corpus-Based Interpreting Studies. Linguistic Insights: studies in language and communication*. Berna: Berna: Peter Lang, 2012.

[15] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ser. ACL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 423–430.

[16] H. Kuramoto, "Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais," *Ciência da Informação*, vol. 25, no. 2, pp. 182–192, 1996.

[17] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[18] T. Chung, "A corpus comparison approach for terminology extraction," *Terminology*, vol. 9, pp. 221–246, 2003.