

Visual characterization of biomedical texts with word entropy

Mirko Degli Esposti
Dipartimento di Matematica
Università di Bologna
desposti@dm.unibo.it

Paolo Rosso
NLEL, ELiRF
Universidad Politécnica de Valencia
proso@dsic.upv.es

Roxana Danger
NLEL, ELiRF
Universidad Politécnica de Valencia
rdanger@dsic.upv.es

Sandra García Blasco
NLEL, ELiRF
Universidad Politécnica de Valencia
sangarbl@epsa.upv.es

Abstract. Recently, the relation between the entropy of words (a new measure from Information Theory introduced by Montemurro in 2001) and the role of words in literary texts, as well as the capacity of entropy for clustering words, has been shown. Our final goal is to investigate if and how the list of ranked words (using entropy) can be useful in other more practical contexts, such as information retrieval task or automatic classification of bio-medical textual data. In this work, we analyze the effectiveness of the keywords selected by the Montemurro's approach to capture the semantics behind biomedical text collections, and using the spectrum of words we offer a visual representation of the text's content. Besides, we compare the resulting keyword lists with the ones obtained with TF-IDF measure, and discuss some of the most interesting facts obtained from this comparison.

1. Introduction

In [1] and [2], Montemurro et al. introduced a novel (i.e. not linguistic) method for keywords extraction and applied to literary texts, such as novels or scientific books. It appears reasonable to investigate if and how this method, combined with the more founded and powerful computational linguistic methods, can be useful in other more practical contexts, such as information retrieval task or automatic classification of bio-medical textual data [3]. This short paper represents a very first contribution to this new line of research.

In this context, we analyze two relatively small corpora of biomedicine: DrugTarget (DT), and Protein-protein interaction (PPI). Both domain areas are quite near semantically, and therefore, they share many terms of vocabulary. Drug targets are the molecular structures whose abnormal activity, associated to a disease, can be modified by drugs, improving the health of patients. Protein interactions allow to interpret how organisms work and this knowledge consequently permits to modify their behavior at bio-molecular and upper levels. However, the discoveries are used by different professionals: pharmaceutics in the DT domain, and biologist in the PPI domain.

In this way, we expect to analyze if the keywords selected by the Montemurro's approach allows to capture the semantics behind text documents. Besides, we compare the resulting keyword lists with the obtained with TF-IDF measure, and we describe some of the most interesting facts obtained from this comparison.

2. Methods

Leaving aside most of the mathematical details in [1] and [2], it is important to recall the main simple and essential ideas beyond the method, also in order to motivate the experiments performed here and discussed in the next section.

First of all, in its initial implementation, any text under analysis is considered just as a sequence of words, with no stop-words eliminated and without any kind of stemming. The basic idea at the heart of the method is extremely simple: the semantic relevance of a given word is not given (only) by its frequency, but mostly by its spatial distribution along the text. More precisely, significant words in a given text tend to be *clustered* in certain sections of the text. (e.g. Fig. 2). Any indicator able to detect this spatial clustering can be used to rank words, and the entropy of words introduced in [1] and in [2] is just one of these indicators.

Roughly speaking, given a text of N words and for a given integer P (to be determined), the text is split in P parts, each with roughly $s = N/P$ words. Given now an arbitrary word ω , appearing n_ω times along the whole texts, a $p_j = p_j(\omega) = n_j/n_\omega$ is defined, where n_j is the number of occurrences of word ω in the j -th piece of the text, with $j = 1, 2, \dots, P$. From these p_j , the traditional Shannon's entropy of the given word ω can be calculated: $\hat{h}(\omega, P) = -\sum_{j=1:P} p_j \log p_j$. Moreover, using simple analytical tools, one can calculate the entropy of the same given word $\hat{h}(\omega, P)$, when *averaged over all possible shuffling* of the original text. The

This work has been partially funded by the project TEXT-ENTERPRISE 2.0 (TIN2009-13391-C04-03) and "Juan de la Cierva" Program, both of the Ministerio de Ciencia y Tecnología, Spain.

significance of the word is then given by $\Delta I(\omega, P) := \hat{h}(\omega, P) - h(\omega, P)$, i.e. by the deviation of its entropy with respect to the entropy of the same word, with the same frequency in a complete random text.

A crucial step is the choice of parameter P , or better the choice of the *characteristic scale* $s = N/P$ to use in the computation of $\Delta I(\omega, P)$. This step can be performed by calculating the *whole information content* of the text: $\Delta I(s) := \sum_{\omega} \mu_{\omega} \Delta I(\omega, P)$, where the sum is over all possible words, and $\mu_{\omega} = n_{\omega} / N$ is the overall probability of the word. Typically, $\Delta I(s)$ has a maximum as a function of $s = N/P$, indicating the scale with maximum information content of words and suggesting the specific value to use in calculating the entropy.

3. Results and Discussion

In the first experiment, for simplicity, we analyze the differences between the vocabularies of PPI and DT corpora and how the information behind words could be used for classification tasks using just positive examples and the title and abstract of texts. DT corpus [3] is composed of 1500 positive articles respectively, with articles published between 1995 and 2001; PPI corpus is composed of 3019 articles published between 1987 and 2008, referenced by IntAct database.

As described in the previous section, keyword selection according to the entropy of words needs the estimation of the characteristic scale of the texts, that is, the scale s that maximize the information content, $\Delta I(s)$. In Fig. 1 the plot of $\Delta I(s)$ as a function of s is shown for DT and PPI corpora, yielding a $P \approx 600$ and $P \approx 1200$ as characteristic scale of the two corpora, respectively.

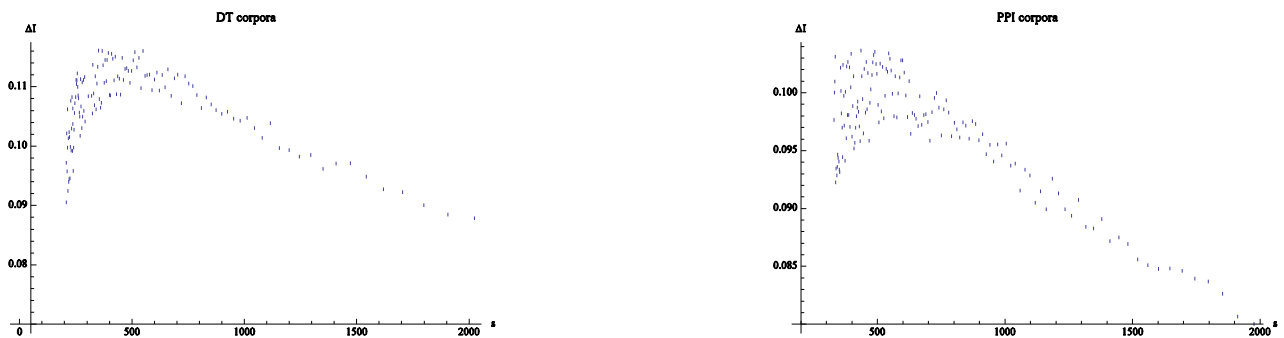


Figure 1: Overall entropy as a function of the scale $s = N/P$ for the *DT corpus* (left) and for the *PPI corpus* (right).

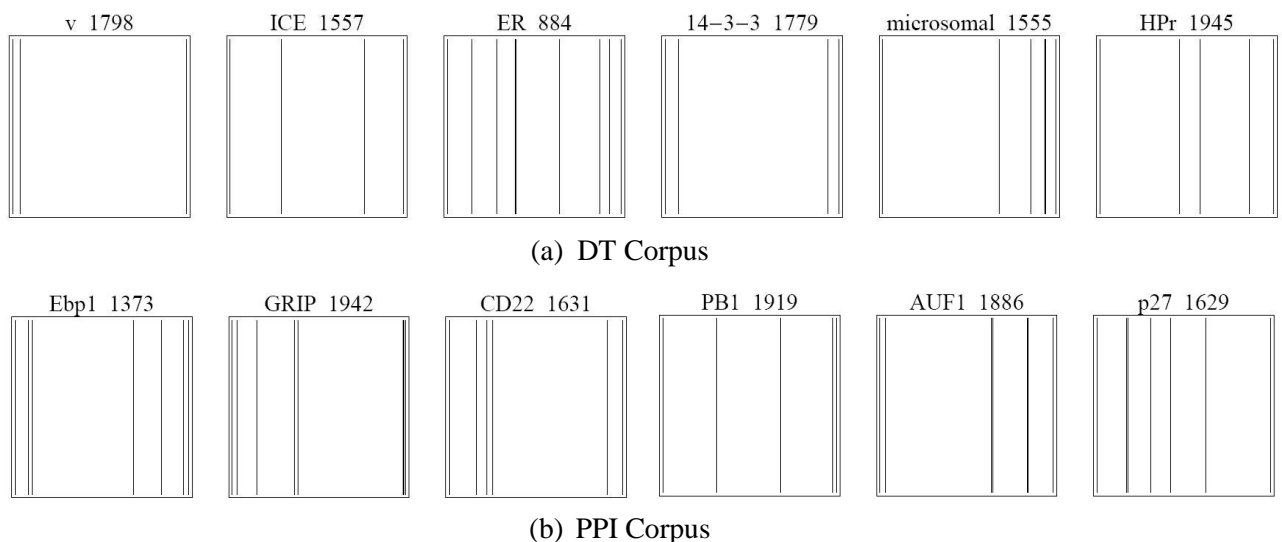


Figure 2: Spectrum of the 6 most entropic words for each corpus: vertical lines correspond to occurrences of the given word along the corpora (x-axis).

For each corpus, we have ranked words with respect to the values of $\Delta I(\omega, P)$, and in both cases, the set of selected words are clearly related with the main topic:

- For PPI corpus, words are generally gene or proteins (or part of such terms), methods used in PPI experiments or cellular processes. Protein interactions are essential for the majority of cellular processes, and these interactions are experimentally demonstrated using standardized methods.

- For DT corpus, the vocabulary is more associated to receptors, kinases, and cell location. The exception here is “chi” (used in some cases as the statistical test). According to their definitions, receptor and protein kinases are good candidates to be drug targets.

In another experiment, we randomly mixed abstracts from the two corpora. In this case, the 100 most entropic words are genes and proteins, which are the common semantics of the two corpora. This supports the hypothesis that entropy is a measure suitable to characterize the domain of a given text collection.

We carried out a further experiment in order to compare the resulting lists with those obtained using TF-IDF (term frequency - inverse document frequency) metric [4]. Each word in a corpus was ranked accordingly to its average of TF-IDF over the texts of the corpus. We relaxed the word measure, in the sense of IDF using a parameter to define a minimal percentage of documents containing the word, $P_{minDocs}$. We have noticed that the semantics of the selected words using TF-IDF is the same of the words selected with entropy. Also, for percentages near to 1% the lists of the 2000 first ranked words for entropy and TF-IDF are very similar, having a 77% of common words. The 50 first words ranked by TD-IDF, using $P_{minDocs}$ allowing the maximum matching with respect to entropy list, are listed in Table 1, for both corpora (common words are marked with *). The percentages of common ranked words for different $P_{minDocs}$ are shown in Table 2.

Table 1: First 50 ranked words according avg. TF-IDF

DT ($P_{minDocs}=1\%$)				PPI ($P_{minDocs}=0,775\%$)			
Words	Avg. TF-IDF	Words	Avg. TF-IDF	Words	Avg. TF-IDF	Words	Avg. TF-IDF
Na	0,0802	potassium*	0,0423	IL	0,1409	TAP	0,0671
GTP*	0,0762	channels*	0,0423	catenin	0,1208	JNK*	0,0671
Rnase*	0,0760	prostate*	0,0420	Bcl	0,1201	integrin*	0,0651
ORFs*	0,0706	galactose	0,0419	p53*	0,1061	TNF*	0,0634
CoA*	0,0592	Ser*	0,0417	Fas*	0,1018	Ca	0,0614
heme*	0,0570	steroid*	0,0413	E2F*	0,0950	iron*	0,0600
nt*	0,0551	trypsin*	0,0411	Cdc42*	0,0909	AP	0,0598
UDP	0,0538	NADH	0,0407	Z	0,0846	cap*	0,0597
hydroxylase*	0,0528	ABC*	0,0403	EF	0,0844	Skp1*	0,0590
tRNA*	0,0527	beta1*	0,0400	Rac*	0,0814	H4*	0,0587
protease*	0,0526	Cys*	0,0399	insulin*	0,0801	TRAF	0,0577
cytochrome*	0,0521	acyl*	0,0397	EGF*	0,0762	SCF*	0,0560
ADP*	0,0511	death	0,0394	delta*	0,0736	centromere*	0,0558
cardiac*	0,0491	subjects*	0,0394	kinetochore*	0,0736	MAPK*	0,0540
synthetase*	0,0480	isomerase*	0,0392	circadian*	0,0731	cAMP	0,0537
L	0,0479	nucleoside	0,0386	meiotic*	0,0723	Golgi*	0,0531
apoptosis*	0,0478	ORF*	0,0386	Wnt*	0,0708	Red	0,0530
NAD	0,0458	pyridoxal*	0,0385	kappaB	0,0705	PDZ*	0,0527
D	0,0452	methylation*	0,0384	Ca2	0,0695	myosin*	0,0517
oxidase*	0,0445	lipid*	0,0379	caspase*	0,0694	14*	0,0516
methyltransferase*	0,0441	C2	0,0373	NF	0,0694	checkpoint*	0,0515
dehydrogenase*	0,0440	sodium*	0,0371	ER*	0,0693	E2*	0,0514
delta*	0,0439	P2	0,0369	phytochrome*	0,0689	heterochromatin*	0,0513
cyclase*	0,0435	cAMP*	0,0367	Ras*	0,0678	channels*	0,0508
Ca2	0,0431	subtilis*	0,0366	prostate*	0,0672	V	0,0508

Table 2: Number of common words between the 2000 first ranked words, comparing entropy and TF-IDF for different values of $P_{minDocs}$.

	0,000%	0,100%	0,325%	0,550%	0,775%	1,000%	2,000%	3,000%	4,000%	5,000%
PPI	20	198	514	473	1553	1432	928	688	573	506
DT	30	108	381	559	1374	1540	1050	815	675	583

References

- [1] Montemurro, M. A. and Zanette, D. H., “Entropic analysis of the role of words in literature texts”. Journal of Advances in Complex Systems (ACS), 5(1):7-17, 2002.
- [2] Montemurro, M. A. and Zanette, D. H., “Towards the quantification of the semantic information encoded in written language”. Journal of Advances in Complex Systems (ACS), 13(02):135-153, 2010.
- [3] Danger R, Segura-Bedmar I, Martinez P, Rosso P., “A comparison of machine learning techniques for detection of drug target articles”. J Biomed Inform, 2010.
- [4] Wu H.C., Luk R.W.P., Wong K.F., Kwok K.L., “Interpreting TF-IDF term weights as making relevance decisions”. ACM Transactions on Information Systems, 26(3):1-37, 2008.