2013 12th Mexican International Conference on Artificial Intelligence

Analysis and Transformation of Textual Energy Distribution

Alejandro Molina*, Juan-Manuel Torres-Moreno*, Eric SanJuan*, Gerardo Sierra† and Julio Rojas-Mora‡
*Université d'Avignon et des Pays de Vaucluse
Laboratoire Informatique d'Avignon
{alejandro.molina-villegas, juan-manuel.torres, eric.sanjuan}@univ-avignon.fr
†Universidad Nacional Autónoma de México
Instituto de Ingeniería
{amolinav, gsierram}@iingen.unam.mx
‡Institute of Statistics
Universidad Austral de Chile
julio.rojas@uach.cl

***Abstract*—In this paper we revisit the Textual Energy model. We deal with the two major disadvantages of the Textual Energy: the asymmetry of the distribution and the unboundedness of the maximum value. Although this model has been successfully used in several NLP tasks like summarization, clustering and sentence compression, no correction of these problems has been proposed until now. Concerning the maximum value, we analyze the computation of Textual Energy matrix and we conclude that energy values are dominated by the lexical richness in quadratic growth of the vocabulary size. Using the Box-Cox transformation, we show empirical evidence that a log transformation could correct both problems.**

***Keywords*-weight model; Sentence Importance; Measures of Informativity;**

I. Introduction

One of the major advances in Natural Language Processing (NLP) is the possibility of processing textual content in numerical ways. Since the Vector Space Model, proposed in [1], it has been possible to represent texts by vectors. Each element of the vector is assigned a value, called weight, which represents the importance of a word in the text. The most widely used technique, TF-IDF [2], assigns a value according to the number of occurrences of a word in a text (Term Frequency) and a numerical statistic which reflects how important a word is to a document in a collection of documents (Inverse Document Frequency). Other weight models have been proposed as well; (see for example *Latent Semantic Indexing* [3] and LexRank [4]), all of them sharing the main advantage of numerical processing: language independence.

However, in some cases, we must face the problem of low frequencies in the vocabulary. This is common when vectors represent sentences, phrases or short texts like tweets and blog opinions. Therefore, we must consider alternative weight models, adapted to low frequency vectors.

The Textual Energy model [5], [6] was conceived from the beginning to work with unitary frequencies. In its original definition, phrases are represented as binary vectors. Nevertheless, this model presents two serious disadvantages studied in this paper.

In section II we present the Textual Energy model. Then, in section III we explain how textual energy values have to be computed from scratch. In section IV, we present an analysis of the upper boundary of textual energy based on computations of section III. Finally, in section V we use the Box-Cox transformation to process one thousand text fragments and we conclude that a logarithmic transformation could correct the textual energy distribution avoiding external parameters.

II. From Ising's Magnetic Model to Textual Energy Model

The Textual Energy model was originally inspired by a physics model. The main idea is that words in a document could be coded as magnetic spins with two possible states: "word present" (represented by the spin +1) and "word absent" (represented by the spin -1). Following this analogy, every word is influenced by the others, directly or indirectly, since they all belong to the same system. Thus, as the spins influence each other in a physical system, the words in the document interact giving sense to other words and coherence to the whole text.

We used as a basis the ideas of the magnetic model proposed by Ising [7]. In Ising's model, the magnetic moment of a material is coded by discrete variables that represent spins; each variable has one of two possible values. It is supposed that the spins interact with each others but they are not influenced by external fields (which is a simplification of the real interactions). Such a system might be represented by a complete graph in which nodes represent spins and arcs represent the interaction between spins. In Fig. 1, the complete graph with eight units (K_8) represents a system in which three units (u_1, u_3 and u_6) have a positive spin and five units (u_2, u_4, u_5, u_7 and u_8) have a negative one.

Hopfield adopted Ising's ideas and applied them to Neural Networks theory [8]. In the Hopfield network, every unit has

978-1-4799-2605-3/13 $31.00 © 2013 IEEE
DOI 10.1109/MICAI.2013.32

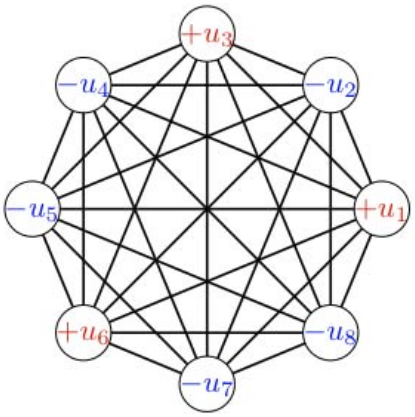

Figure 1. A complete graph with eight units K_8 representing a Hopfield network.

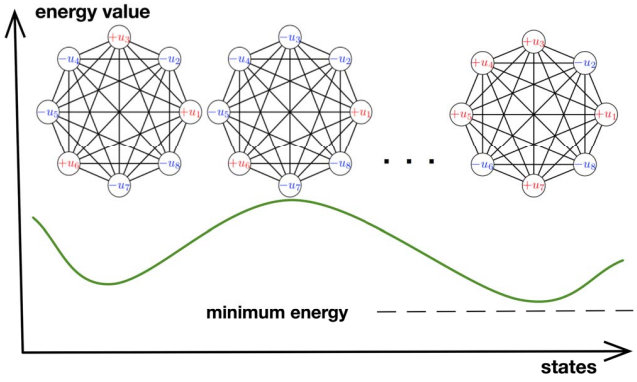

Figure 2. Example of energy changes in a Hopfield network.

an activation state (in practice a binary variable) and units are fully interconnected as in the Ising model. The *connection strength* between two units u_i and u_j is described by a real number representing the weight value: $\mathbf{weight}(u_i, u_j)$. The network has the following restrictions:

- There are no auto-connections: $\mathbf{weight}(u_i, u_i) = 0 \ \forall i$;
- Connections are symmetric: $\mathbf{weight}(u_i, u_j) = \mathbf{weight}(u_j, u_i) \ \forall i, j$.

One of the main applications of the Hopfield network is the *associative memory*. The main idea is that the network memorizes input patterns constantly repeated in the coded input examples. Some of the input might contain noisy information that the network would try to ignore. In the training process, spins are adapted depending on the current input. This adaptation of spins modifies weights according to Hebb's rule described in [9]. The result is a different network state at each iteration and an associated value of the state. At this stage, the network could change drastically giving the impression of an electric discharge. That is the reason why Hopfield called this value "the energy". When the system stops modifying weights it has converged to a stable state of low energy. This allows the recognition of learned patterns. Fig. 2 depicts the example of the network in Fig. 1 converging to a minimum energy state.

One issue of the associative memory is that it could quickly be saturated. Thus, only 14% of the patterns are memorized correctly. This situation has limited its practical applications, as discussed in [9]. Nevertheless, several authors adapted this model to NLP and proposed the Textual Energy showing good results in topic segmentation[6], summarization [10], clustering [11] and sentence compression [12].

III. The Textual Energy computation

Let T be the number of unique terms of a document, the size of the vocabulary. We can represent a phrase φ as a string of T spins where the term i could be present or absent. Thus, a document with Φ phrases, is composed of Φ vectors in a space of dimensions $[\Phi \times \mathrm{T}]$ represented by a textual matrix.

Let (1) be the document-term matrix of a document with Φ phrases and T terms.

$$\mathbb{A}_{[\Phi \times \mathrm{T}]} = \begin{pmatrix} a_{1,1} & \cdots & a_{1,\mathrm{T}} \\ \vdots & \ddots & \vdots \\ a_{\Phi,1} & \cdots & a_{\Phi,\mathrm{T}} \end{pmatrix} \tag{1}$$

In $\mathbb{A}$, $a_{i,j} = \mathbf{tf}(\varphi_i, w_j)$ is the frequency of the term w_j in the phrase φ_i.

As proposed in [5], [10], the interaction weights between terms are computed using Hebb's rule in its matricial form by the product (2).

$$\mathbb{J} = \mathbb{A}^{\mathrm{t}}\mathbb{A} \tag{2}$$

where $j_{i,j} \in \mathbb{J}$ describes the frequencies of terms occurring in the same phrase and terms co-occurring with previously related terms.

The Textual Energy matrix (3) is computed by two successive products. The energy values, $e_{i,j} \in \mathbb{E}$, represent the degree of relationship between φ_i and φ_j.

$$\mathbb{E} = \mathbb{A}\mathbb{J}\mathbb{A}^{\mathrm{t}} = (\mathbb{A}\mathbb{A}^{\mathrm{t}})^2 \tag{3}$$

In [12], the following two major disadvantages of the Textual Energy were found when authors tried to apply it to measure the informativeness of intra-phrase discourse segments.

Observation 1: Textual Energy values have no upper bound, $e_{i,j} \in [0, \infty)$, because the maximum energy value of a text depends entirely on word local frequencies (integer values). The consequence is the difficulty for comparing documents through their Textual Energy values.

Observation 2: The distribution of Textual Energy values is skewed and biased towards low values; perhaps following

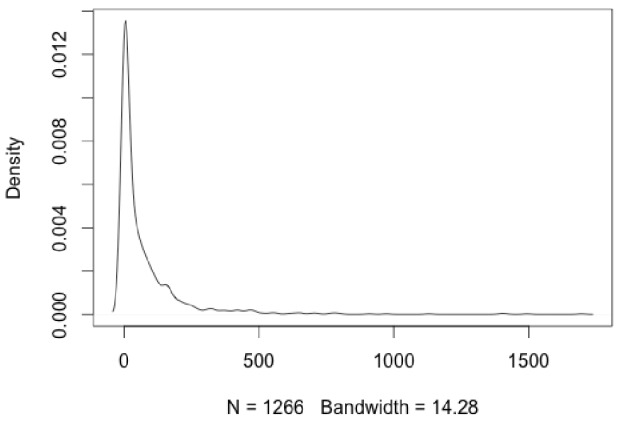

Figure 3. Density plot of Textual Energy distribution for 1,266 discourse segments.

a power law distribution. As a consequence, differences among low energy values are not significant enough to distinguish important phrases from unimportant ones.

Fig. 3 shows the density plot of the distribution of more than one thousand discourse segments of the corpus of sentence compression in Spanish[1]. We observe highly asymmetric characteristics of Textual Energy values. In section IV we theoretically analyze **Observation 1** and in section V we propose alternatives to **Observation 2**.

IV. Analysis of the upper bound of Textual Energy

In order to find the upper bound of Textual Energy, we first consider the easiest case, when the document-term matrix $\mathbb{A}$ contains only binary values. Under these conditions, Textual Energy values are maximal if all the terms appear in all the phrases of the document, $a_{i,j} = 1; \quad \forall a_{i,j} \in \mathbb{A}$.

$$\mathbb{A}_{[\Phi \times \mathrm{T}]} = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}. \tag{4}$$

According to section III, we must compute $(\mathbb{A}\mathbb{A}^{\mathrm{t}})$ where every element in the resulting matrix will be computed as $\sum_{i=1}^{\mathrm{T}}(1 \times 1)$:

$$(\mathbb{A}\mathbb{A}^{\mathrm{t}}) = \begin{pmatrix} \underbrace{1^2 + \cdots + 1^2}_{\mathrm{T}} & \cdots & \underbrace{1^2 + \cdots + 1^2}_{\mathrm{T}} \\ \vdots & \ddots & \vdots \\ \underbrace{1^2 + \cdots + 1^2}_{\mathrm{T}} & \cdots & \underbrace{1^2 + \cdots + 1^2}_{\mathrm{T}} \end{pmatrix}. \tag{5}$$

[1]Corpus available at http://molina.talne.eu/sentence_compression/data.

By (3), the Textual Energy values are those of matrix $(\mathbb{A}\mathbb{A}^{\mathrm{t}})^2$ where every element will be computed as $\sum_{i=1}^{\Phi}(\mathrm{T} \times \mathrm{T})$ resulting (6).

$$\begin{aligned} \mathbb{E} &= \begin{pmatrix} \underbrace{\mathrm{T}^2 + \cdots + \mathrm{T}^2}_{\Phi} & \cdots & \underbrace{\mathrm{T}^2 + \cdots + \mathrm{T}^2}_{\Phi} \\ \vdots & \ddots & \vdots \\ \underbrace{\mathrm{T}^2 + \cdots + \mathrm{T}^2}_{\Phi} & \cdots & \underbrace{\mathrm{T}^2 + \cdots + \mathrm{T}^2}_{\Phi} \end{pmatrix} \\ &= \begin{pmatrix} \Phi\mathrm{T}^2 & \cdots & \Phi\mathrm{T}^2 \\ \vdots & \ddots & \vdots \\ \Phi\mathrm{T}^2 & \cdots & \Phi\mathrm{T}^2 \end{pmatrix} \end{aligned} \tag{6}$$

Consequently, Textual Energy values $e_{i,j}$ are $O(\Phi\mathrm{T}^2)$ for the binary case. However, the number of phrases is always less than the number of terms ($\phi \leq \mathrm{T}$). So we can say that $e_{i,j}$ are $O(\mathrm{T}^2)$.

In the general case where frequencies could be greater than one, let us consider a constant c to be the maximum frequency value and suppose that all term frequencies are equal to this value; $\mathbf{tf}(\varphi_i, w_j) = c, \quad \forall i \in [1, \Phi], j \in [1, \mathrm{T}]$. By analogy to (5), elements in $(\mathbb{A}\mathbb{A}^{\mathrm{t}})$ can be computed as $\sum_{i=1}^{\mathrm{T}}(c \times c)$ and according to (6), the elements of $(\mathbb{A}\mathbb{A}^{\mathrm{t}})^2$ will take the form $\sum_{i=1}^{\Phi}(\mathrm{T}c^2 \times \mathrm{T}c^2) = \underbrace{\mathrm{T}^2c^4 + \cdots + \mathrm{T}^2c^4}_{\Phi \text{ times}}$.

It turns out that, in general, Textual Energy values $e_{i,j}$ are $O(\Phi\mathrm{T}^2c^4)$. However, we argue that it is always T^2 which dominates the growth. Although the maximum frequency c has the highest exponent, in practical situations, it is the size of the vocabulary that determines the final energy values. In real documents, the maximum term frequency in the same sentence is $1 \leq c \leq 3$. In fact, in our experiments with one thousand discourse segments $1 \leq c \leq 2$ always holds. Taking this into account, we can affirm that Textual Energy values are $O(k\mathrm{T}^2)$ where $k = \Phi c^4$ could be considered a constant.

This result means that lexical richness of a document determines the maximum energy value. For instance, given an artificial document with 1 000 word types (one page and a half) we could reach Textual Energy values of 10^6 assuming a great lexical richness.

V. Normalization of the Textual Energy using the Box-Cox transformation

The Box-Cox transformation is typically used in asymmetric distributions like that of Textual Energy. This transformation is a continuous function with a single parameter λ which could be adjusted to correct the distribution [13] .

In the case of the Textual Energy, $e_{i,j}$ values are transformed in $e_{i,j}^{\lambda}$ according to (7),

$$e_i^{\lambda} = \begin{cases} K_1(e_i^{\lambda-1}) & \text{if } \lambda \neq 0, \\ K_2\mathbf{log}(e_i) & \text{if } \lambda = 0 \end{cases} \tag{7}$$

where K_2 corresponds to the geometric mean (8) and K_1 depends on λ and K_2 according to (9).

$$K_2 = \left(\prod_{i=1}^{n} e_i \right)^{1/n} \tag{8}$$

$$K_1 = \frac{1}{\lambda K_2^{\lambda-1}}. \tag{9}$$

After iterating parameter λ from -1 to 1 with a step of 0.001 we found that the best parameterization, the nearest to a normal distribution, is obtained for $\lambda = -0.015$. Since the best parametrization is obtained when $\lambda \approx 0$, we argue that textual energy can be corrected simply by applying the logarithm function. As suggested in [13], in this situation external parameters could be avoided for adjusting the distribution of the values. Therefore, we propose (10) to measure the energy (the informativeness) of a sentence φ_i taking into account the whole document context.

$$\textbf{Info}(\varphi_i, \Phi) = \frac{\textbf{log}(\textbf{E}(\varphi_i) + 1)}{\text{argmax}_{\varphi_i \in \Phi}(\textbf{log}(\textbf{E}(\varphi_i) + 1))}. \tag{10}$$

This last score has a quasi-normal distribution and is $[0, 1]$ bounded.

The density plot after transformation using (10) for the same 1 266 discourse segments is shown in Fig. 4. We observe that although there are still some problems for values between 0 and 0.4, the curve looks smoother and well distributed than that obtained using pure energy values in Fig. 3. Advantages of the proposed transformation are more clear when we observe the application to a real text.

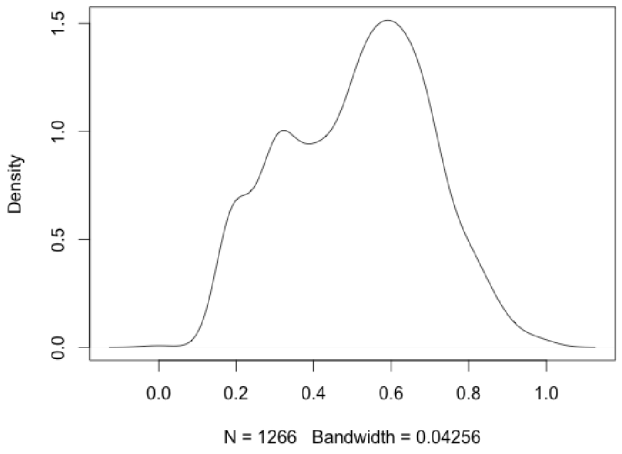

Figure 4. Density plot of transformed Textual Energy distribution for 1 266 discourse segments.

Fig. 5 shows an example of the effect of transforming Textual Energy values using a simple normalization (only dividing by the maximum value) and using the score of (10). The first column corresponds to the simple normalization and the second one to the score. Grey tonalities were mapped from real values in the $[0, 1]$ interval. Dark gray means low values (near to 0) and light gray is used for high values (near to 1). The dashed line in the middle corresponds to the global mean value and the bars in the cells represent the deviation from this value. We observe that after the log transformation it would be possible to assign textual segments to two well balanced classes, low energy and high energy, by comparing deviations. Segments near the mean value would be considered of normal importance and segments above mean value would be considered as the highly important ones. Besides, the sensitivity of original Textual Energy has been increased for less important segments. Before transformation, none of the segment of the second sentence "*Their survival depends on their ability to regulate the expression of genes coding for the enzymes and transport proteins, required for growth in the altered environment*" is considered as important. In fact, only its second segment is considered non-zero.

VI. Conclusions

We have improved the Textual Energy model which has been used for scoring terms and phrases in Summarization, Topic Segmentation, Text Classification, Sentence Compression and other tasks. We have analyzed the reasons of its two major disadvantages: asymmetry of the distribution and unboundedness of the maximum value. We have proved that the maximum value of Textual Energy is determined by the lexical richness of a document. In future experiments, we will analyze this phenomenon empirically, using different sizes of document corpora: tweets, Wikipedia documents and scientific articles, in order to verify our hypothesis on short, medium and long documents.

After applying the Box-Cox transformation to over one thousand discourse segments we have found that the optimal value of parameter λ is very close to zero. According to the Box-Cox theoretical framework, we suggest to correct the Textual Energy distribution using a log transformation score.

We conclude that Textual Energy must be corrected before its use for comparison proposes and the easiest way of doing it, without external parametrization, is using a log transformation.

ACKNOWLEDGMENT

We would like to thank Peter Peinl and Patricia Velazquez for all their help and support. This work was partially supported by the CONACyT grant 211963 and project 178248.

REFERENCES

[1] G. Salton, *The SMART Retrieval System – Experiments un Automatic Document Processing*. Prentice-Hall, Englewood Cliffs, 1971.

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

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[4] G. Erkan and D. R. Radev, "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization," *Journal of Artificial Intelligence Research*, vol. 22, no. 1, pp. 457–479, 2004.

[5] S. Fernández, E. SanJuan, and J.-M. Torres-Moreno, "Energie textuelle des mémoires associatives," in *Proceedings of Traitement Automatique de la Langue Naturelle (TALN'07)*, vol. 1, Toulouse, France, 5-8 Juin 2007, pp. 25–34.

[6] S. Fernández, E. SanJuan, and J.-M. Torres-Moreno, "Textual energy of associative memories: performants applications of enertex algorithm in text summarization and topic segmentation," in *Mexican International Conference on Artificial Intelligence (MICAI'07)*, Aguascalientes, Mexico, 2007, pp. 861–871.

[7] D. J. Amit, H. Gutfreund, and H. Sompolinsky, "Statistical mechanics of neural networks near saturation," *Annals of Physics*, vol. 173, no. 1, pp. 30–67, 1987.

[8] J. J. Hopfield and D. W. Tank, ""neural" computation of decisions in optimization problems," *Biological cybernetics*, vol. 52, no. 3, pp. 141–152, 1985.

[9] J. A. Hertz, A. S. Krogh, and R. G. Palmer, *Introduction to the theory of neural computation*. Westview press, 1991, vol. 1.

[10] S. Fernández, E. SanJuan, and J.-M. Torres-Moreno, "Enertex : un système basé sur l'énergie textuelle," in *Proceedings of Traitement Automatique des Langues Naturelles (TALN'08)*, Avignon, France, 2008, pp. 99–108.

[11] A. Molina, G. Sierra, and J.-M. Torres-Moreno, "La energía textual como medida de distancia en agrupamiento de definiciones," in *Journées d'Analyse Statistique de Documents (JADT'10)*, Rome, 2010.

[12] A. Molina, J.-M. Torres-Moreno, E. SanJuan, I. da Cunha, and G. S. Martínez, "Discursive sentence compression," in *Computational Linguistics and Intelligent Text Processing*. Springer, 2013, pp. 394–407.

[13] G. Box and D. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 211–252, 1964.

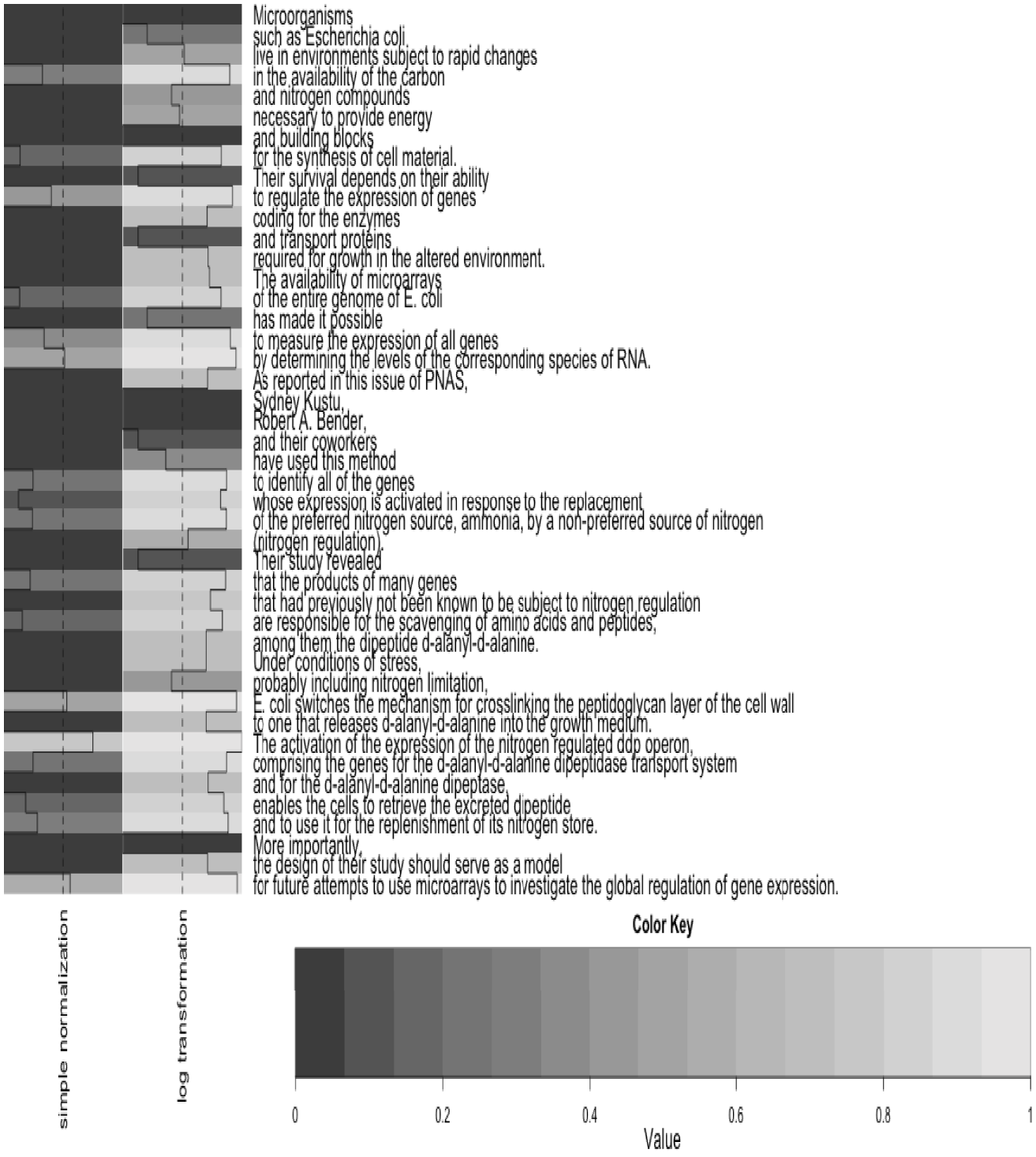

Figure 5. Example of two transformations of textual energy values for discourse segments in a scientific abstract.