Refining character relationships using embeddings of textual units

Guillaume Guex[1]

[1]Departement of Language and Information Sciences, University of Lausanne, Switzerland

Abstract
A clear and well-documented LaTeX document is presented as an article formatted for publication by CEUR-WS in a conference proceedings. Based on the "ceurart" document class, this article presents and explains many of the common variations, as well as many of the formatting elements an author may use in the preparation of the documentation of their work.

Keywords
LaTeX class, paper template, paper formatting, CEUR-WS

1. Introduction

Distant reading tools allow researchers, from various fields, to quickly gain knowledge on textual corpora without actually reading them. Purposes of these methods are various, but can be mainly categorized into two groups: in the first case, these methods are used in order to tag, classify, or summary large quantities of texts, in order to quickly structure information or to deliver a speech over the whole studied corpus. Methods in this case rely heavily on Big Data and make an extensive use of Machine Learning algorithms. In the second case, researchers use these methods to underline hidden structures in a particular text, helping them to refine their understanding of it and reinforce stated hypotheses. Methods in this setting can also rely on Machine Learning, but must typically be build with more caution and attention to details: corpus are smaller, analyses are closer to the work, and methods must be more transparent in order to appropriately interpret results.

Automatic extraction and analysis of *character networks* from literacy works typically belong in the latter group. These methods aim at representing various interactions occurring between fictional characters found in a textual narrative with a graph, thus showing explicitly the hidden structure of character relationships constructed by the author. This structure might allow to find hidden patterns within a book, which can highlight a particular genre or style.

COMHUM 2022: Workshop on Computational Methods in the Humanities, June 9–10, 2022, Lausanne, Switzerland
guillaume.guex@unil.ch (G. Guex)
0000-0003-1001-9525 (G. Guex)

2. Methodology

When building character networks from a textual narrative, the most widespread method consists in dividing the studied work into n textual units $u_1, \dots, u_n$, which can be, e.g., sentences, paragraphs, or chapters, and then counting characters co-occurrences in these units. Usually, the text constituting these units is discarded and the resulting network displays edges which roughly represent an aggregated number of interactions between characters. However, by doing so, the aggregation occurs on various type of interactions and will give little information about the type of relationship which exist between characters. In this paper, we propose a data organization leading to various methods which takes into consideration the text contained in the unit, helping to refine understanding of characters and their relationships.

2.1. Data organization

In this article, the data representation for a textual narrative, divided in n textual units $u_1, \dots, u_n$, is made through two tables. The first one is well known in the field of textual analysis, and consists in the $(n \times v)$ contingency table $\mathbf{N}$, as represented by Table 2.1, where v is the vocabulary size. In this table, each row represents an unit, each column a word, and cells n_{ij} counts the number of times the word i appears in the unit. Using this table typically denotes a *Bag-of-Words* approach in our analyses.

The second table, denoted by $\mathbf{O}$, has a size of $(n \times p)$ where p is the number of *character-based objects* found in the narrative and cells o_{ij} counts occurrences of object j in unit i. A character-based object can be loosely defined in order to be flexible for various types of narrative or analyses, but can be roughly seen as a countable recurring narrative entity, containing one or more characters. For example, it can be a character occurrence, a character co-occurrence, an oriented character interaction (e.g.

	aller	allumer	apercevoir	bas	bon	...
u_{100}	23	2	6	11	6	...
u_{7795}	12	1	0	3	9	...
u_{7796}	10	0	5	1	5	...
u_{7797}	0	0	1	0	0	...

Table 1

A snippet of the contingency table $\mathbf{N}$ extracted from *les Misérables*. Rows are chapters, columns are words in the vocabulary, and cell n_{ij} counts the number of time word j appear in chapter i.

a dialog), or even a particular recurring event containing multiple characters (e.g. a lecture). In this article, we mostly consider character occurrences and pair of characters co-occurrences, as shown in Table 2.1.

	Cosette	Thénardier	Valjean	...
u_{101}	66	78	0	...
u_{102}	21	26	0	...
u_{103}	12	25	0	...
u_{104}	12	0	5	...

	...	Cosette-Thénardier	Cosette-Valjean	...
u_{101}	...	66	0	...
u_{102}	...	21	0	...
u_{103}	...	12	0	...
u_{104}	...	0	5	...

Table 2

A snippet of the character-based objects table $\mathbf{O}$ extracted from *les Misérables*. Rows are chapters, columns are character occurrences (top) and character co-occurrences (bottom), and cell o_{ij} counts the number of times object j appear in chapter i. Character co-occurrences are computed here as $o_{ij} = \min(o_{ik}, o_{il})$, where $\{k, l\}$ are characters constituting j.

This data organization already gives an orientation to the subsequent analyses and should be kept in mind of the practitioner. Textual unit are now considered as *individuals* (in the statistical terminology), defined by their *variables* contained in the different columns of both tables. Moreover, subsequent analyses are oriented in searching how the object table $\mathbf{O}$ has an influence over the contingency table $\mathbf{N}$, i.e. searching which words are over-represented or under-represented knowing the character-based objects in the unit. While authors use characters in order to build the narrative, we, to a certain extent, work backward: we are searching how character appearances and interactions in the textual unit act on her/his choice of words. If extraction methods permit it, a practitioner should include all character-based objects which she/he desire to study. Here, for example, the choice to include character co-occurrences along with occurrences is motivated by the fact that we are interested in studying character relationships. A character co-occurrence can roughly be seen as an interaction between two characters, and an interaction between two characters (or more, but higher order interactions are outside the scope of this article) should be considered as an object of its own: this interaction is not necessarily the sum of its parts and gives a particular flavor to the unit.

This data organization also highlight the importance of choosing a proper size for the units. These units should be large enough to contain enough words in order the properly capture the textual specificity of each unit, but not too large, as each unit should ideally captures particularities about one of the character-based objects. Unfortunately, it is impossible to define an ideal size for all types of analysis and this size should be balanced regarding the level of analysis, text size, selected character-based objects, and previous knowledge of the studied work.

The use of a contingency table to represent the textual resource present in the units denotes a *Bag-of-Words* approach. Using this approach loses the information relative to the order of words in the units, but permits to transform a chain of characters, improper to statistical analyses, into a contingency table, a well studied mathematical object which permits the use of various kind of statistical methods. The next section shows two methods of analysis based on this table.

2.2. Embedding of textual units

Various methods can be performed on the contingency table $\mathbf{N}$ in order to extract information from it, such as Sentiment Analysis (REF), Textometry (REF), or even Deep Learning methods. Here, we make to choice to extract a lower dimensional, numeric representation of each units, in other words, an *embedding of textual units*. In section (REF), these vectors of textual units are used as anchor points in order to also embed character-based objects in the same space. Therefore, it is crucial that an interpretation about the directions, or the regions, of this embedding is possible, in order to properly extract information about the localization of character-based objects vectors (the relative position of these vectors is insufficient). For that reason, we focus on embeddings of textual units which also contain *vectors of words*: by examining the positions of character-based objects relatively to word vectors, these objects can be characterized.

We propose here two embeddings verifying this condition: Section (REF) describes *Correspondence Analysis (CA)* and section (REF) focuses on *Pre-trained word vectors (WV)*.

2.2.1. Correspondence Analysis (CA)

Using Correspondence Analysis in order to analyze textual resources has a long tradition in literature (REF). It has the advantage to naturally provide an embedding space, the factorial map, where units are placed alongside word vectors, allowing to analyze or represent the different units in terms of word specificities. Moreover, units and words placement in the embedding space have a direct interpretation in terms of chi2 distance between profiles, which is desirable when interpreting results.

By performing a Correspondence Analysis on table $\mathbf{N}$, we get n vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ corresponding to units (rows) and v vectors $\mathbf{w}_1, \ldots, \mathbf{w}_v$ corresponding to words (columns). Each of these vectors has a size of $\min(n, v) - 1$, which will generally be $n - 1$. For detailed computations of quantities given by the CA, see Appendix A.1. The position of a unit vector $\mathbf{x}_i$ can be interpreted by two different methods.

The first methods is to look at its position on a particular axis α, i.e. looking at the component $x_{i\alpha}$. This component gives the position of the unit on a particular found "contrast" between units, depending on the chosen α, and this contrast will be strong if p_α, i.e. *the proportion of inertia expressed in* α, is high. It can further be interpreted by observing the *contribution of each word* j composing the axis α, i.e. $c^w_{j\alpha}$ and looking at the sign of $w_{j\alpha}$. Hopefully, by extracting and interpreting the most positive and negative contributing words, the chosen axis can express a particular duality which operates inside the studied text (e.g. love-hatred, action-description, lightness-darkness). The second method is to directly look at the *similarities* s_{ij} between an unit i and a word j, as defined by the scalar product between their vectors $s_{ij} := \mathbf{x}_i^\top \mathbf{w}_j$. A positive (resp. negative) similarity denotes a over-representation (resp. under-representation) of the word j in i, which permit to find lists of words characterizing the different units. Observe that in this article, both methods are rather applied to character-based object vectors defined in section (REF), as they lay in the same space as unit vectors.

Note that vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ obtained from CA reflect textual unit profile (in terms of words) regarding the mean profile. This mean that this analysis is in fact *contrastive*: we highlight unit variations among the studied text. This could therefore becomes problematic if we would like to build our dataset on multiple works: units belonging to the same text have a greater chance to be close from each other, and the displayed variance might be largely composed by differences of style between text. Moreover, the words helping the interpretation of units (and character-based objects) are contained in the studied text, which might limit its range of application.

2.2.2. Pre-trained word vector

Pre-trained word vectors, based on Word2Vec (REF), Glove (REF), or Fasttext (REF), have recently received great attention from various fields (REFS). The generally originate from a training of very large corpora, such as Wikipedia (REF) or Common Crawl (REF) and results in a embedding containing a large quantities of word vectors.

2.2.3. Character and character pairs embeddings

3. Results

4. Conclusion

A. Appendix

A.1. Correspondence Analysis

Starting from the $(n \times v)$ contingency table $\mathbf{N} = (n_{ij})$, we define the vector of unit weights as $\mathbf{f} = (f_i) := (n_{i\bullet}/n_{\bullet\bullet})$ and the vector of word weights as $\mathbf{g} = (g_j) := (n_{\bullet j}/n_{\bullet\bullet})$, where $\bullet$ denotes the summation on the replaced index. It is then possible to compute the *weighted scalar product matrix between units* $\mathbf{K} = (k_{ij})$ with

$$k_{ij} := \sqrt{f_i f_j} \sum_{k=1}^{v} g_k (q_{ik} - 1)(q_{jk} - 1), \qquad (1)$$

where $q_{ik} = \frac{n_{ik} n_{\bullet\bullet}}{n_{j\bullet} n_{\bullet k}}$ is the *quotient of independence* of the cell i, k. The vector of textual unit i, $\mathbf{x}_i = (x_{i\alpha})$, is obtained by the eigendecomposition of the matrix $\mathbf{K} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ and with

$$x_{i\alpha} := \frac{\sqrt{\lambda_\alpha}}{\sqrt{f_i}} u_{i\alpha}, \qquad (2)$$

where λ_α are the eigenvalues contained in the diagonal of $\mathbf{\Lambda}$ and $u_{i\alpha}$ the eigenvectors components found in $\mathbf{U}$. We can also find the vector of word j, $\mathbf{w}_j = (w_{j\alpha})$, with

$$w_{j\alpha} := \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^{v} f_i q_{ij} x_{i\alpha}. \qquad (3)$$

Note that various other quantities of interest can also be computed in CA, such as

$$p_\alpha := \frac{\lambda_\alpha}{\lambda_\bullet} \text{ : the proportion of inertia expressed in } \alpha,$$

$$c^u_{i\alpha} := \frac{f_i x^2_{i\alpha}}{\lambda_\alpha} \text{ : the contribution of unit } i \text{ to axis } \alpha,$$

$$c^w_{j\alpha} := \frac{g_j w^2_{j\alpha}}{\lambda_\alpha} \text{ : the contribution of word } j \text{ to axis } \alpha,$$

$$h^u_{i\alpha} := \frac{x^2_{i\alpha}}{\sum_\alpha x^2_{i\alpha}} \text{ : the contribution of axis } \alpha \text{ to unit } i,$$

$$h^w_{j\alpha} := \frac{w^2_{j\alpha}}{\sum_\alpha w^2_{j\alpha}} \text{ : the contribution of axis } \alpha \text{ to word } j,$$

For a detailed interpretation of these different quantities, see (REF).

References