Machine Assisted Study of Writers' Rewriting Processes

Julien Bourdaillet[1], Jean-Gabriel Ganascia[1], and Irène Fenoglio[2]

[1] University Pierre and Marie Curie - LIP6,
104 Quai Kennedy, 75016 Paris, France
{Julien.Bourdaillet, Jean-Gabriel.Ganascia}@lip6.fr
[2] Ecole Normale Supérieure - ITEM,
45 rue d'Ulm, 75005 Paris, France
Irene.Fenoglio@ens.fr

Abstract. This paper presents a joint work between artificial intelligence and literary studies. As part of the humanities, *textual genetic criticism* deals with writers' rewriting processes. By studying drafts and manuscripts issued from these processes, the genesis of the text is discovered. When draft comparison is done manually, it requires a huge amount of work. The introduction of the machine provides a high gain on efficiency and enables to focus on the interpretative work. The application we developed relies on a sequence alignment algorithm close to the ones used in molecular biology. This paper describes the textual alignment algorithm, presents an experimental validation, and illustrates the textual analysis with two genetic studies.

1 Introduction

Textual genetic criticism is a school of literary studies born during the seventies in France [1]. Its main contribution is to study all the versions of a text, the final edition becoming one version among others. This broadening of the scope of the study introduces a temporal dimension by dealing with a sequence of versions. Finally, the goal is to make the genesis of the text emerge from the drafts in order to understand the writers' rewriting processes.

Two draft examples are presented in Figure 1. From these manuscripts, linearized transcriptions are extracted, resulting in a sequence of versions. The comparison of these versions enables the identification of invariants and differences between them. When the author inserts or deletes a sentence or a paragraph, the comparison is not so hard; but when spelling faults are corrected or micro-modifications are done, this becomes tedious. Further, authors work on their style, they sometimes want to "free" a word or a phrase in order to use it at another position in the text, that is to move a word or a phrase. This move detection is very hard for humans, but it is a requirement for genetic interpretation because moves are one of the four operations defined by genetic criticism, along with insertions, deletions and replacements.

In fact, this corresponds to the notion of *edit distance with moves* which has been studied in theoretical computer science [2]. This model is an extension of the usual edit distance which can be computed in a quadratic time; but the introduction of the

move operation makes the problem NP-complete and there exists no polynomial time algorithm [3].

It means that the automation of this textual comparison work is a difficult problem for computer scientists. In this paper we present our application named MEDITE dedicated to this work [4], an experimental validation, and two genetic studies.

2 Algorithm

In this section we present our algorithm named MEDITE; it is closely related to fragment alignment commonly used in bioinformatics [5]. The two texts are considered as two character sequences A and B which are processed in five steps. The first step is a pre-processing step where character equivalence classes are set between identical upper- and lower-case characters, accentuated or not characters and separators. Geneticists choose which equivalence classes to use depending of the analysis type.

The second step identifies repeated character blocks between A and B by building a generalized suffix tree between A and B [6]. This data structure enables to find all repeated character blocks between A and B. The size of this set of blocks is exponential and only a subset is interesting: the subset of super maximal exact matches (SMEMs), that is matches of maximal size which are not included in other SMEMs.

The third step aligns these SMEMs in order to determine which are invariant and which are moved. The pairwise alignment of SMEMs enables to make the decision: aligned SMEMs are considered as invariant and unaligned as moved. Because the space of possible alignments is combinatorial, we use an A^* heuristic algorithm based on the computation of the symmetric difference between remaining block to align.

The fourth step consists in looping over the pairings resulting from step 3 and in considering the subsequences in A and B between each pair of aligned blocks. These subsequences are processed again with steps 2 and 3. It allows the pairing of new blocks which are then included in the main alignment. This recursive step enables the pairing of blocks which would otherwise have been left unaligned.

The last step is the deduction of insertions, deletions and replacements. Insertions, deletions and replacements can then be deduced. Deletions are non-repeated blocks in A and insertions are non-repeated blocks in B. Further, when there is a deleted block d in A and an inserted block i in B between two pairs of aligned blocks, and the ratio $|d|/|i|$ reaches a threshold t, then d and i are transformed in replacements r_1 and r_2, meaning that r_1 has been replaced by r_2. t is arbitrarily set to 0.5.

3 Experimental Validation with Synthetic Data

In this section we present an experimental validation of the algorithm. The goal is to evaluate the quality of MEDITE on synthetic data when a reference alignment exists. Given a text and a noise model, a second text is generated by altering the first one; the alignment between the texts is recorded during the alteration process. Then, it is possible to evaluate the quality of the algorithm by comparing its results with the reference alignment. This kind of experiment is not possible with texts issued from genetic criticism because there exists no reference alignment.

This experiment compares MEDITE with GREEDY which is a greedy algorithm for computing the edit distance with moves [3]. GREEDY pairs the longest blocks first and considers them as moves, finally it computes the classical edit distance by dynamic programming. Five modified documents were generated with the noise model from a 20 KB text. Then the documents were aligned with MEDITE and accuracy rates evaluated. Two series of tests with different modification ratios were conducted: in the first one there are 5 % of insertions, 5 % of deletions, 5 % of replacements, and 5 % of moves, which means that there are 20 % differences between original and modified texts; in the second series, the ratio is set to 10 %, meaning that there are 40 % differences. For each of the five kinds of characters (insertions, deletions, replacements, moves and invariants) the accuracy rate is defined as the number of correctly aligned characters / the total number of characters. Then the average accuracy rate is calculated for these five rates. For the average weighted accuracy rate, the four accuracy rates are weighted with their ratio of the texts' sizes. For example, for the first test series, we have $weighted\ accuracy = 0.80 * Invariant\ acc. + 0.05 * Insertion\ acc. + 0.05 * Deletion\ acc. + 0.05 * Replacement\ acc. + 0.05 * Move\ acc.$. The average runtime of alignments is also calculated. Figure 1 presents the results of this experiment.

Table 1. Synthetic data alignment with GREEDY and MEDITE

Algorithm	GREEDY		MEDITE	
Modification ratio	5 %	10 %	5 %	10 %
Average accuracy	39.20 %	37.72 %	84.87 %	80.19 %
Av. weighted accuracy	75.16 %	59.75 %	93.45 %	86.26 %
Average runtime	19mn 25s	48mn 59s	0mn 9s	0mn 14s

For the average precision, GREEDY's results are around 40 points inferior to MEDITE's results, and for the average weighted precision around 25 points inferior. Further, the runtimes are an order of magnitude different.

4 Machine Assisted Genetic Analysis

4.1 Short Text Analysis

We present a genetic analysis of a short passage of a novel named "La Robe Noire" (The Black Dress), by Andrée Chedid a Franco-Egyptian contemporary writer, where she describes her relations with her mother when she was a teenager. Figure 1(a) presents the manuscript of which several linearized transcriptions have been extracted: they are the *versions* of the text. These versions are compared automatically using MEDITE, then the following manual interpretations are done.

In this text, most of the modifications are related to the mother character "Elle" (She). The main fact discovered during this study is the complete transformation of the writer's viewpoint to her mother, between the first and the last version.

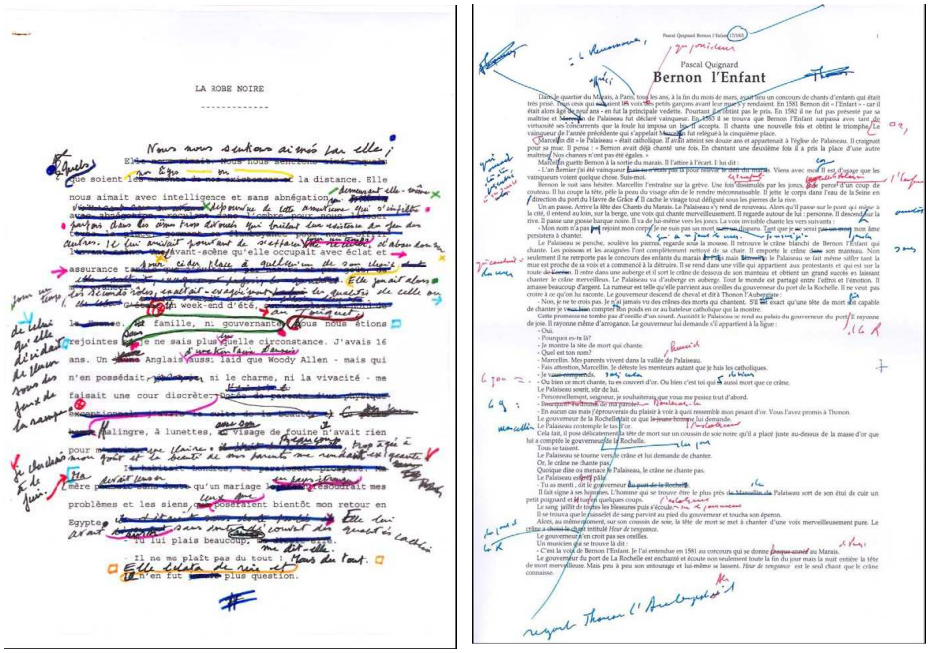

Fig. 1. Manuscript exemples: (a) “La Robe Noire” by Andrée Chedid, and (b) “Fête des Chants du Marais” by Pascal Quignard

The first version describes the mother as a loving and maternal mother. It starts by the sentence “Elle nous aimait.” (She loved us.). But the set of modifications transforms her in an imposing and brilliant character. The first sentence strength is reduced becoming “Nous nous sentions aimés par elle” (We felt loved by her). In the first version, the phrase “l’avant scène qu’elle occupait avec éclat et assurance” (the first place she was occupying with brightness and assurance) was clearly minimized by the preceding passage “gommant sa flamboyance pour nous offrir l’avant-scène.” (minimizing her flamboyant style to offer us the first place). In the last version, the minimization is restricted: “ Il lui arrivait pourtant de s’effacer pour un temps, de reculer, d’abandonner” (Sometimes, she kept in the background for a while, leaving). The phrase “parfois avec abnégation” (sometimes with self-abnegation) becomes “avec intelligence et sans abnégation” (with intelligence and without self-abnegation). The following sentence is inserted: “Elle jouait alors pour un temps, les seconds rôles, exaltait - exagérément parfois les qualités de celle ou de celui qu’elle décidait de placer sous les feux de la rampe.” (For a while she was playing the second roles, exalting, sometimes excessively, the qualities of the one she decided to give the first place.); it clearly shows that she controls and manipulates the social positions inside the family.

This short text shows the interest of genetic analysis. The meaning of the first and the last versions is completely different and can not be guessed by studying only the last

edited version. MEDITE enables to make easier the comparison work and to discover this kind of genetic evolution.

4.2 Genetic Folder Analysis - Global Viewpoint of a Text Genesis

We present a genetic analysis with the help of MEDITE of a tale named “Fête des Chants du Marais” (Swamp Song Feast) by Pascal Quignard, a French contemporary writer. The tale relates a song contest for children in the Marais (a Paris district) during the sixteenth century. The author has produced five manuscripts (or versions) of the text and each version has been annotated giving different sub-versions, called *states*; this gives 15 different states of the text. The set of these states is called the *genetic folder*.

The early analysis consists in comparing the first and the last state of the text, the following facts can be observed:

- The rewriting process is expansive; there are more insertions than deletions.
- By examining replacements, we remark that the words “Palaiseau” (main character’s first name) and “gouverneur” (governor character) were added several times. These additions reinforce the weight of those characters in the text and focus the attention on them.
- Moves involve only single words, there are no phrase moves.

This first analysis enables to browse modifications and to discover interesting ones. It is then possible to look for the *genetic moment* of a modification. For example, if we wonder when the following sentence appears: “On monte le crâne au Grenier des Réformés” (The skull is carried up to the Reformeds’ Attic), by comparing the versions, we observe that between the second and the third version the sentence “On monte le crâne au Grenier” is added, and between the third and the fourth version “des Réformés” is added. This enables to observe the genesis of specific text portions and to understand that the final version results from multiple rewritings.

More generally, we analysed the 15 states with MEDITE by comparing the states pairwise and chronologically; three *rewriting campaigns* have been discovered:

1. Between the first and the second version: The main characteristic of this campaign is to focus on the style; the author improves his style by bringing the text closer to the linguistic norm.
   For example, the following modifications have been done: “du visage” (of the face) becomes “de son visage” (of his face); “soit” becomes “fût”, present subjunctive to imperfect subjunctive of to be; “- Pourquoi ? Tu doutes” (- Why ? You doubt) becomes “Le gouverneur: - Douterais-tu” (The governor: - Would you doubt), moving from oral style to written style.
2. Between the second and the third version: This campaign focuses on the beauty theme. A lot of terms related to beauty are inserted.
   For example, the following expressions are added: “qui était très beau” (who was very beautiful); “le Beau Palaiseau” (the Beautiful Palaiseau); “Son bel air s’ajoute à sa beauté” (His beautiful attitude improves his beauty”); “radieux, le bel adolescent” (shining, the beautiful teenager).

Further, some expressions are magnified: “à l’église” (to the church) becomes “à la chantrerie de la basilique” (to the basilica’s chant-school); “le crâne” (the skull) becomes “la tête de mort” (the death’s head).

3. Between the third and the fifth version, the campaign focuses on passion. The following expressions are added: “passionément catholique” (passionately catholic); “Le Palaiseau souffre et pleure.” (The Palaiseau suffers and cries.); “Je les déteste” (I hate them); “Nous ne l’aimons guère” (We dislike it).
   Further, by considering the positions of these expressions, we can remark that they enable to strengthen the distinction and the opposition between Catholics and Reformeds which is latent in the text.

By comparing the last state of the first version and the first state of the second version, we discovered that there are a lot of modifications. They were typed by the writer: MEDITE makes appear these modifications that are not visible on the manuscript (these ones where there are no pen corrections). They correspond to the first stylistic rewriting campaign. On the other hand, for the third campaign on passion, corrections are not typed: the author corrects directly on the manuscript, with his pen, because more considerations are needed to emphasize the focus on passion.

5 Conclusion

In this paper, MEDITE has been presented; it addresses the problematic of textual alignment issued from genetic criticism. The algorithm is based on character sequence alignment using a suffix tree and an A* heuristic alignment algorithm. It handles fine-grained modification detection and move identification. The experimental validation on synthetic data shows good results. Two genetic analyses were presented; they showed how the machine helps the geneticist in the study of rewriting processes. MEDITE is now used by text geneticists for literary studies.

References

1. Deppman, J., Ferrer, D., Groden, M., eds.: Genetic Criticism - Texts and Avant-textes. University of Pennsylvania Press (2004)
2. Lopresti, D.P., Tomkins, A.: Block Edit Models for Approximate String Matching. Theoretical Computer Science **181** (1997) 159–179
3. Shapira, D., Storer, J.A.: Edit Distance with Move Operations. In: CPM. Volume 2373 of Lecture Notes in Computer Science., Springer (2002) 85–98
4. Bourdaillet, J., Ganascia, J.G.: Alignment of Noisy Unstructured Text Data. In: Proc. of the IJCAI Workshop on Analytics for Noisy Unstructured Text Data (AND 2007) of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007). (2007) pp. 139–146
5. Bray, N., Dubchak, I., Pachter, L.: AVID: A Global Alignment Program. Genome Res. **13** (2003) 97–102
6. Gusfield, D.: Algorithms on Strings, Trees and Sequences: Computer Science and Computer Biology. Cambridge University Press (1997)