Exploratory Text Segmentation through Joint Distribution Estimation

Dominique Gay[*], Romain Guigourès[**]
Marc Boullé[*],[***] Fabrice Clérot[***]

[*]Université de La Réunion
[**]Zalando
[***]Orange Labs

Résumé. We suggest a novel way for exploratory topic segmentation based on data grid models. In this context, a text can be represented as a data set of two-dimensional points; each point is defined by two variables: a word (categorical value) and the placement of the word in the text (numerical value). Instantiating data grid models to the 2D-points turns the problem into coclustering. Simultaneously, the words are partitioned into clusters and the placement (or time) variable is discretized into intervals/segments, following a parameter-free Bayesian model selection approach. We also suggest several criteria for exploiting the resulting grid through agglomerative hierarchies, for interpreting the clusters of words and characterizing their components through insightful visualizations. Experiments on the Bible show the relevance of our approach.

1 Exploratory topic segmentation

Text Segmentation has been extensively studied over the past years since it is a prequel to further text analytics like e.g., text summarization or information retrieval. Since pioneering work Hearst (1997) based on lexical cohesion, many text segmentation techniques have been suggested in the literature (see Purver (2011) for a well-structured survey) : e.g, among others, Utiyama et Isahara (2001), MCSeg Malioutov et Barzilay (2006), BayesSeg Eisenstein et Barzilay (2008), HierBayes Eisenstein (2009), APS Kazantseva et Szpakowicz (2011), TSM Du et al. (2013), etc.

In this paper, we focus on the long text segmentation problem and we suggest a new approach for *exploratory topic segmentation*. Pragmatically, a resulting text segmentation should hold the following features :

- *(i)* the segmentation technique should give a global picture of the underlying structure of the text and show the evolution of the detected topics along the running text
- *(ii)* the segmentation technique should highlight groups of words that are characteristic of segments ;
- *(iii)* for the sake of ergonomy, computing the segmentation should not involve parameter tuning ;
- *(iv)* the whole methodology should allow to explore the resulting segmentation at various granularities.

To the best of our knowledge, there is no long text segmentation technique having all these features. The methodology we suggest fulfills all the previous requirements and uses recent progress in joint distribution estimation based on data grid models Guigourès et al. (2015); Gay et al. (2015).

2 Data Grid Models in a Nutshell

Data grid models Boullé (2011) aim at estimating the joint distribution between K variables of mixed-types (categorical as well as numerical). The main principle is to simultaneously partition the values taken by the variables into groups/clusters of categories for categorical variables and into intervals for numerical variables. In this context, a text is represented by two variables : W for the words and T for the placement (or time) of the word in the text. Instantiating data grid models for text segmentation, the result is a two-dimensional data grid whose cells (or co-clusters) are defined by a part of each partitioned variable value set, i.e., a cluster of words and a time/placement interval.

In order to choose the "best" data grid model M^* (given the data) from the model space $\mathcal{M}$, we use a Bayesian Maximum A Posteriori (MAP) approach. We explore the model space while minimizing a Bayesian criterion, called cost. The cost criterion implements a trade-off between the accuracy and the robustness of the model and is defined as follows :

$$cost(M) = -\log(\underbrace{p(M \mid D)}_{\text{posterior}}) \propto -\log(\underbrace{p(M)}_{\text{prior}} \times \underbrace{p(D \mid M)}_{\text{likelihood}}) \tag{1}$$

Thus, the optimal grid M^* is the most probable one (maximum a posteriori) given the data. Considering a data-dependent hierarchical prior (on the parameters of the grid model) that is uniform at each stage of the hierarchy, Boullé (2011) has shown that we can obtain an exact analytical expression of the cost criterion. The full details about the *cost* criterion and the optimization algorithm are available in Boullé (2011). The key features to keep in mind are : *(i)* the algorithm is parameter-free, i.e., there is no need for setting the number of clusters/intervals per dimension ; *(ii)* using a greedy bottom-up strategy coupled with pre and post-optimization heuristics and Variable Neighbourhood Search meta-heuristic, it provides an effective locally-optimal solution to the data grid model construction efficiently, in sub-quadratic time complexity ($O(N\sqrt{N}\log N)$ where N is the number of data points). Notice that in the case of two categorical variables (e.g., texts × words for text categorization), the criterion is an exact density estimation estimator, that asymptotically converges to the mutual information between both partitions. In other words, it could be compared as a regularized version of the Information Theoretic Coclustering Dhillon et al. (2003) – while, in addition it can deal with mixed-typed variables.

2.1 Data grid exploitation and visualization

When facing long texts, the optimal grid M^* can be made of hundreds of parts per dimension, i.e., many thousands of cells, which is difficult to exploit and interpret. To alleviate this issue, we suggest a grid simplification method together with several criteria that allow us to choose the granularity of the grid for further analysis, to rank words in clusters and to gain insights in the underlying text through meaningful visualizations.

Dissimilarity index and grid structure simplification. In order to simplify the structure, we propose to apply an agglomerative hierarchical clustering on top of the optimal model M^*. We derive a dissimilarity measure between clusters adjacent intervals from the exact criterion described above.

Definition 1 (Dissimilarity index) *Let $c_{.1}$ and $c_{.2}$ be two parts of a variable partition of a grid model M. Let $M_{c_{.1} \cup c_{.2}}$ be the grid after merging $c_{.1}$ and $c_{.2}$. The dissimilarity $\Delta(c_{.1}, c_{.2})$ between the two parts $c_{.1}$ and $c_{.2}$ is defined as the difference of cost before and after the merge :*

$$\Delta(c_{.1}, c_{.2}) = cost(M_{c_{.1} \cup c_{.2}}) - cost(M) \tag{2}$$

Building such a hierarchy on top of a grid obtained using a regularized approach has the advantage to provide an interesting exploratory analysis tool for exploring the results at different level of granularity, while ensuring reliable results. It has been shown, in Guigourès et al. (2015),that asymptotically, Δ converges to the Jensen-Shannon divergence between the distributions of two clusters over the time intervals (resp. two intervals over the clusters). This approach could then be regarded as an agglomerative information bottleneck Slonim et Tishby (2000) approach starting from an optimal level and preventing the probability estimation errors occurring at the first merges of agglomerative approaches based on divergence. In order to control the degradation of the quality of our model, we introduce the information ratio of the grid M', defined as follows :

$$IR(M') = \frac{cost(M') - cost(M_\emptyset)}{cost(M^*) - cost(M_\emptyset)} \tag{3}$$

where $M_\emptyset$ is the null model (the grid with a single cell).

Typicality for ranking words in a cluster. When the grid is coarsen during the hierarchical agglomerative process, the number of clusters of words decreases and the number of words per cluster increases. It is useful to focus on the most representative words among thousands of words of a cluster. In order to rank words in a cluster, we define the typicality of a word as follows.

Intuitively, the typicality evaluates the average impact in terms of *cost* on the grid model quality of removing a word from its cluster and reassigning it to another cluster. Thus, a word is representative (say typical) if it is "close" to the cluster it belongs to and "different in average" from other clusters. The typicality is actually very similar to the *Silhouettes* approach Rousseeuw (1987), an efficient technique for validation and interpretation of a clustering.

Insightful visualizations with Mutual Information. It is common to visualize 2D coclustering results using 2D frequency matrix or heat map. We also suggest an insightful measure for co-clusters to be visualized, namely, the Contribution to Mutual Information (CMI) – providing additional valuable visual information inaccessible with only frequency representation.

Definition 2 (Contribution to mutual information) *The mutual information between two partitioned variables W^M of size J_W and T^M of size J_T (from the partition M of W and T*

variables induced by the grid model M) is defined as :

$$MI(W^M;T^M) = \sum_{i_1=1}^{J_W}\sum_{i_2=1}^{J_T} MI_{i_1i_2} \text{ where } MI_{i_1i_2} = p(c_{i_1i_2})\log\frac{p(c_{i_1i_2})}{p(c_{i_1.})p(c_{.i_2})} \tag{4}$$

where $MI_{i_1i_2}$ represent the contribution of cell $c_{i_1i_2}$ to the mutual information, $p(c_{i_1i_2})$ is the observed joint probability of points in cell $c_{i_1i_2}$ and $p(c_{i_1.})p(c_{.i_2})$ is the expected probability in case of independence, i.e., the product of marginal probabilities.

3 Application

Data. For our experiments, we use the text of The Bible which is among the best-selling books of all time – and as far as we know, surprisingly, it has never been automatically segmented. It is also one of the most studied book. Thus, the following findings through exploratory text segmentation can easily be asserted by common knowledge on the book – which is taken as ground truth. The wide-spread King James version of The Bible (without apocrypha) is composed of 66 books [1] ramified in *(i)* the *Old Testament*, subdivided into the Pentateuch (5 books), the Historical Books (12), the Poetical Books (5), the Prophets (17) and *(ii)* the *New Testament*, subdivided into the Gospels (4), the Acts of the Apostles (1), the Pauline and other Epistles (21) and the Revelation (1). Each book is divided into chapters, then each chapter into verses. The Bible originally contains $|W| = 12918$ unique words ($N = 789628$ total) and $|T| = 31102$ verses. For this experiment, we work at the verse level (i.e., cut points are allowed only between verses) and pre-process the text by removing stop words and grouping lexical items by stemming Porter (1980). Thus, the 2D input for data grid models is like : $(1, begin), (1, god), (1, creat), (1, heaven), \ldots,$
$(|T|, lord), (|T|, jesus), (|T|, christ), (|T|, amen)$.

The big picture. Thanks to Khiops Coclustering [2], an effective locally-optimal grid is obtained in 67 minutes and is made of 329 segments and 252 clusters of words. At this scale (see Figure 1), the analysis of the summary provided by the 2D-segmentation is not an easy task for a non-expert. However, we can highlight two clusters of words : In green (rectangular), the cluster of words whose most typical word is "*begat*", which relates to the genealogy of some characters and is a recurrent topic in the Bible (see circles) : e.g., Adam and Noah in the Genesis book, Saul in the Chronicles and Jesus in the Gospels. In pink, the clusters of words whose most typical words are "*angel, repent, satan and throne*" which relates to the apocalypse described in the Revelation book at the end of the Bible.

Zoom into the Gospels. Now, we zoom into the Gospels part (see Figure 2). The storyline of Gospels is Jesus' life. Whereas the gospels are perfectly separated from other books, segmentations *between* Matthew, Mark, Luke and John show a mismatch of about 1-3 verses. However, the recurrent topics of the various Gospels are well-identified. In blue (rectangular), cluster of words $C1$ relates to typical characters of the Gospels : "Jesus, discipl, Peter, John, Simon" who are recurrent along the text ; while $C2$ relates to the various acts and encounters

[1] `https://en.wikipedia.org/wiki/List_of_books_of_the_King_James_Version`

[2] Khiops CoClustering is available as a shareware for research purpose at `http://www.khiops.com`

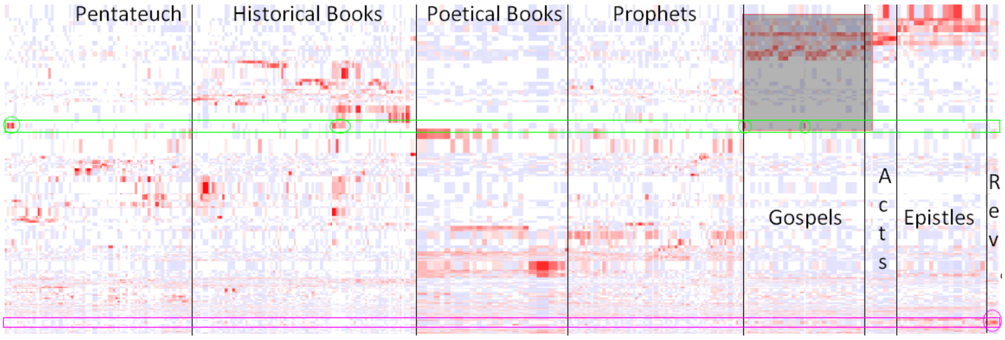

FIG. 1: Visualisation of CMI for the resulting (x-axis = 329 segments × y-axis = 252 clusters of words)-grid obtained on the whole Bible. Red cells indicate positive CMI, i.e., excess of interactions between T and W in the cell. Meta-segmentations in black lines are manually added and annotated.

of Jesus. Typical words of $C2$ are "*ask, temple, sit, pharise, whosoever, ship, heal*". We also observe a strong similitude between the so-called *synoptic* Gospels (Matthew, Mark and Luke) contrasting with the Gospel of John : indeed, considering cells in $C2$, the cumulative CMI in each synoptic Gospel is above 25.10^{-4} while only 5.10^{-4} for the fourth Gospel. Notice also that the genealogy of Jesus is reported only by Matthew and Luke (see green ellipses). Despite the variations between Gospels, they all have in common the passion and resurrection (with typical words : "*pilat, mari, crucifi, sepulchr, betray, . . .* ", placed at the end of each gospel (see blue ellipses).

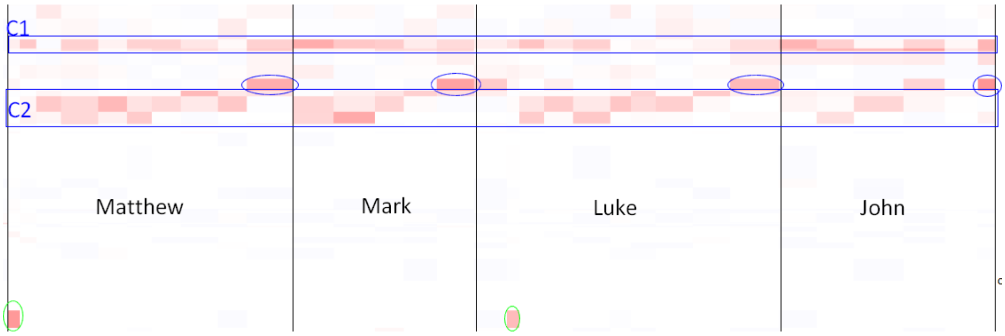

FIG. 2: Zoom into the Gospels.

Agglomerative hierarchy. Since the Bible is made of 66 books, we build a agglomerative hierarchy on top of the computed grid, as defined in previous section, in order to have 66 segments. The resulting segmentation matches perfectly with 13 books boundaries (i.e., matching the end of the books of Ruth, Kings2, Nehemiah, Esther, Ecclesiastes, Song of Solomon, Jeremiah, Ezekiel, Daniel, Zephaniah, Malachi, Gospel of John and Philemon). With a tolerance of ±5 verses, the end of the books of Genesis, Exodus, Leviticus, Job, Psalms, Gospel of Luke,

Acts of the Apostles and the third Epistle of John are also detected. At this granularity, the book of Genesis is still *over-segmented*, since they relate several different stories from the origins of the earth and human kind to the life of characters such as Abraham, Jacob and Joseph – involving a specific vocabulary in each story which is quite rare in the rest of the Bible. Similar observations stand for the book of Exodus – explaining the mismatches.
Continuing towards the null model (see figure 3), the two last steps of segments agglomeration highlight perfect cuts (at the top of the hierarchy) between the Old and the New testament and inside the Old Testament between {Pentateuch, Historical Books} and {Poetic Books, Prophets}.

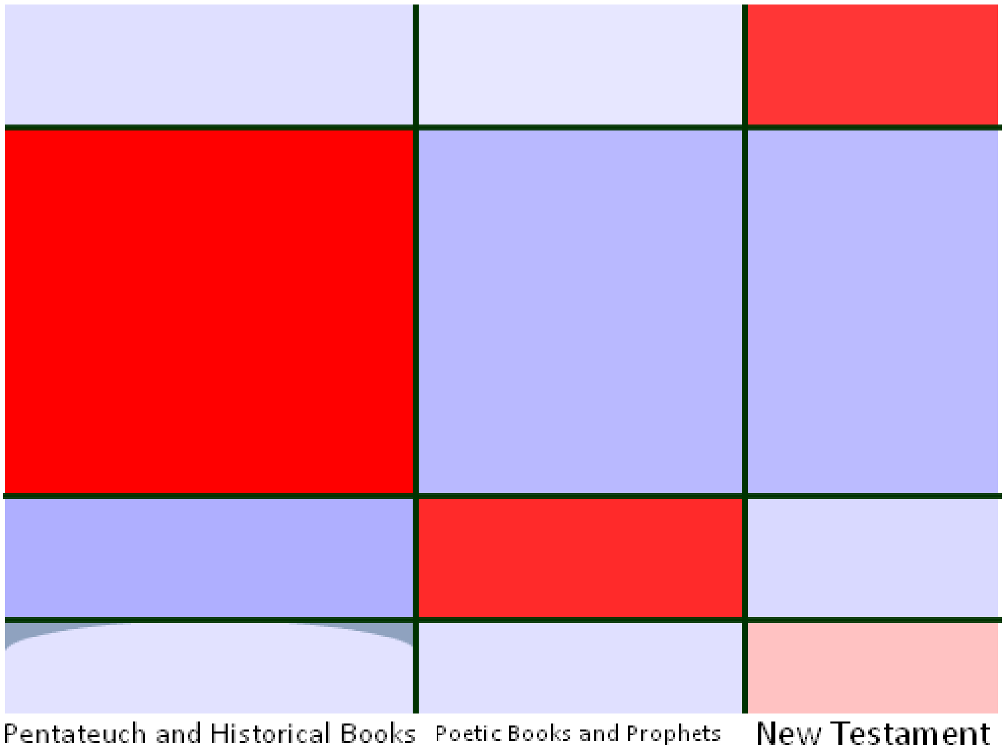

FIG. 3: (3 × 4)-grid, where the New and Old Testament are perfectly separated and where the Old Testament is perfectly divided into {Pentateuch, Historical Books} and {Poetic Books, Prophets}.

4 Discussion

We have suggested a relevant application nugget of data grid models for exploratory topic segmentation. Data grid models provide an effective 2D segmentation of a given long text. The method allows to efficiently get the big picture of the underlying text and to explore the segmentation at multiple levels of granularity while highlighting significant topics of the text.

While applying the method on the Bible has been successful, we plan to extend the experiments to multiple comparisons with state-of-the-art text segmentation techniques on artificially generated data and on other long texts.

Références

Boullé, M. (2011). Data grid models for preparation and modeling in supervised learning. In I. Guyon, G. Cawley, G. Dror, et A. Saffari (Eds.), *Hands-On Pattern Recognition : Challenges in Machine Learning, vol. 1*, pp. 99–130. Microtome.

Dhillon, I. S., S. Mallela, et D. S. Modha (2003). Information-theoretic co-clustering. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003*, pp. 89–98.

Du, L., W. L. Buntine, et M. Johnson (2013). Topic segmentation with a structured topic model. In *Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pp. 190–200.

Eisenstein, J. (2009). Hierarchical text segmentation from multi-scale lexical cohesion. In *Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31 - June 5, 2009, Boulder, Colorado, USA*, pp. 353–361.

Eisenstein, J. et R. Barzilay (2008). Bayesian unsupervised topic segmentation. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 334–343.

Gay, D., R. Guigourès, M. Boullé, et F. Clérot (2015). TESS : temporal event sequence summarization. In *2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, Campus des Cordeliers, Paris, France, October 19-21, 2015*, pp. 1–10.

Guigourès, R., M. Boullé, et F. Rossi (2015). Discovering patterns in time-varying graphs : a triclustering approach. *Advances in Data Analysis and Classification*, 1–28.

Guigourès, R., D. Gay, M. Boullé, F. Clérot, et F. Rossi (2015). Country-scale exploratory analysis of call detail records through the lens of data grid models. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part III*, pp. 37–52.

Hearst, M. A. (1997). TextTiling : segmenting text into multi-paragraph subtopic passages. *Computational Linguistics 23*(1).

Kazantseva, A. et S. Szpakowicz (2011). Linear text segmentation using affinity propagation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 284–293.

Malioutov, I. et R. Barzilay (2006). Minimum cut model for spoken lecture segmentation. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual*

Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program 14*(3).

Purver, M. (2011). Topic segmentation. In G. Tur et R. de Mori (Eds.), *Spoken Language Understanding : Systems for Extracting Semantic Information from Speech*, pp. 291–317. Wiley.

Rousseeuw, P. J. (1987). Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics 20*, 53–65.

Slonim, N. et N. Tishby (2000). Document clustering using word clusters via the information bottleneck method. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 208–215. ACM.

Utiyama, M. et H. Isahara (2001). A statistical model for domain-independent text segmentation. In *Association for Computational Linguistic, 39th Annual Meeting and 10th Conference of the European Chapter, Proceedings of the Conference, July 9-11, 2001, Toulouse, France.*, pp. 491–498.

Summary

We suggest a novel way for exploratory topic segmentation based on data grid models. In this context, a text can be represented as a data set of two-dimensional points; each point is defined by two variables: a word (categorical value) and the placement of the word in the text (numerical value). Instantiating data grid models to the 2D-points turns the problem into co-clustering. Simultaneously, the words are partitioned into clusters and the placement (or time) variable is discretized into intervals/segments, following a parameter-free Bayesian model selection approach. We also suggest several criteria for exploiting the resulting grid through agglomerative hierarchies, for interpreting the clusters of words and characterizing their components through insightful visualizations. Experiments on the Bible show the relevance of our approach.