

Exploratory Text Analysis using Lexical Episode Plots

Valentin Gold, Christian Rohrdantz and Mennatallah El-Assady

University of Konstanz, Germany

Abstract

In this paper, we present Lexical Episode Plots, a novel automated text-mining and visual analytics approach for exploratory text analysis. In particular, we first describe an algorithm for automatically annotating text regions to examine prominent themes within natural language texts. The algorithm is based on lexical chaining to find spans of text in which the frequency of a term is significantly higher than its average in the document. In a second step we present an interactive visualization supporting the exploration and interpretation of Lexical Episodes. The visualization links higher-level thematic structures with content-level details. The methodological capabilities of our approach are illustrated by analyzing the televised US presidential election debates.

Categories and Subject Descriptors (according to ACM CCS): I.5.4 [Pattern Recognition]: Applications—Text processing

1. Introduction

The exploratory analysis of textual data has become a main field in many scientific disciplines. This is partly due to a massive increase in the availability of textual data. Simultaneously, computer-assisted methods for the exploratory analysis of textual data have become increasingly common, user-friendly, and generalized. Hence, computer-assisted applications for text-mining and visualization have become an essential part of most sciences. Yet, existing methods for exploratory text analysis have mainly focused either on illustrating higher-level thematic structures or on close examinations on the content-level. Our approach combines both levels allowing analysts to discover thematic patterns and to derive hypotheses based on a detailed inspection of the content. Hence, we combine the logic of what Digital Humanities scholars call "distant reading" with the logic of "close reading". This approach supports analysts to derive hypotheses such as: What is the thematic structure of the text? Where do textual dynamics come from?

Our contribution is twofold: We first introduce a novel text mining method to identify thematic clusters within a text, named lexical episodes. Second, we propose an interactive visualization to convey the extracted lexical episodes to the analyst enabling a more targeted and efficient analysis of the textual content.

2. Related Work

Since we combine a novel text-mining algorithm with an interactive visualization, we first describe related methods for text extraction and structuring that share some fundamental technical concepts with our approach. Second, we give a short overview on related approaches with respect to text visualizations.

2.1. Segment-based Content Extraction

Our Lexical Episode approach has mainly been inspired by a time series analysis technique [RHD*12]. This technique identifies temporal bursts of individual keywords in textual time series. Keywords are clustered if they are unexpectedly close in time. Moreover, our approach is also related to the concept of *lexical chains*. This concept was first introduced by Morris and Hirst [MH91] and later adapted by others, e.g. [GMFLJ03, Hea97, Hea94]. Lexical chains are segments of related words and indicate units of text being "about the same thing". The original concept as well as the refinements have in common that the distance between the related words to be clustered in a chain/segment is defined arbitrarily by specifying a global threshold based on a fixed number of sentences.

In contrast to these approaches, the lexical episodes relaxes most assumptions. In particular, for the identification of text segments through keyword chains, an individual

threshold is computed based on the distribution of the keyword's index positions in the text. Moreover, our approach indicates the statistical significance of the text segmentation. Finally, each text segment is later refined in a post-process step of pruning.

2.2. Text Visualization

Within the field of text visualization, numerous approaches have been proposed. While there are many tools to visually explore the content of texts and its links to some meta-data, e.g. [O'C14, HCJ14, MH13, AGCH11], most text visualizations apply topic modeling as to convey thematic dynamics over time and space and/or in context of other data facets, e.g. [KJWE14, DWS*12, DWCR11, LZP*09]. However, most approaches either do not provide a thematic overview or do not readily support a close inspection of the textual content. Our approach allows an analysis of arbitrary texts at different levels-of-detail in one visual display. A semantic zooming capability readily enables to swap between close and distant reading and text analysis.

Our approach is also different, as it is based on comprehensible algorithms that allow less experienced analysts to grasp the logic of the lexical episodes. While Topic Modeling is a prominent example for a mostly reliable and useful method for the extraction of thematic descriptors from text to structure its content, the interpretation of their results is often challenging for analysts from the humanities and other related disciplines. Topic Models are usually based on non-deterministic black box algorithms; it is hard for less experienced analysts to properly judge the reliability of results and link them to concrete instances in the textual data (cf. [Sch12, CGW*09]). Yet, the intuitiveness and tangibility of methods applied in research are crucial for humanists.

3. Lexical Episodes

Lexical Episodes are portions within the word sequence of texts where a certain keyword appears more densely than expected from its frequency in the whole word sequence. To find these keywords we focus on the gap in the index positions between two occurrences of the same keyword, i.e. the number of different words in between. Figure 1 illustrates the basic principle. We do not limit the lexical episodes to word unigrams, but also to word ngrams. Since both algorithms are rather similar, we only describe the procedure for word unigrams in more detail.

Formally, a lexical episode for a single morphological lemma l is defined as a sequence of word occurrences w extracted from textual data d that meets several conditions:

$$\text{lexical_episode}(l) = (w_1, w_2, \dots, w_n) \quad (1)$$

$$\text{with } n \geq 3 \quad (2)$$

$$\text{and } l = \text{lemma}(w_1) = \text{lemma}(w_2) = \dots = \text{lemma}(w_n) \quad (3)$$

$$\text{and } i_x = \text{index_position}(w_x, d) \quad (4)$$

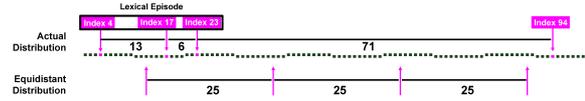


Figure 1: Illustration of the basic principle to define a Lexical Episode. The 100 dots represent a sequence of 100 words. Within this sequence, a certain keyword occurs at the index position 4, 17, 23, and 94. Since the gap between these keywords is smaller than the expected gap (25 words), a Lexical Episode is defined for the first three occurrences of the keyword—but not for the last occurrence of the keyword.

$$\text{and } i_{x+1} > i_x \quad (5)$$

$$\text{and } (i_{x+1} - i_x) < \text{threshold}(l) \quad (6)$$

$$\text{with } \text{threshold}(l) = \text{length}(d) \div \text{freq}(l, d) \quad (7)$$

Each of these conditions can be rephrased as follows:

1. A lexical episode consists of a sequence of word occurrences.
2. It must contain at least 3 word occurrences.
3. The morphological lemma of all word occurrences must be the same.
4. i_x is the index position of a word occurrence w_x within the textual data d .
5. The word occurrences of the lexical episode are ordered by increasing index position.
6. Any pairwise gap in index positions between two consecutive word occurrences of one lexical episode is smaller than a certain threshold that depends on the current lemma.
7. This threshold is obtained by dividing the overall amount of word occurrences in the text by the number of occurrences of words with the given lemma in the whole text.

To generate the Lexical Episodes, three consecutive steps are applied. First, for each word –unigrams and ngrams– in the textual data it is analyzed whether the word forms part of one or more Lexical Episodes. Second, a likelihood ratio test is applied to calculate the unexpectedness/significance of each previously defined Lexical Episode. Hence, our algorithm allows to differentiate between various degrees of significance. In cases where unigram and ngrams contain the same lemma, we keep the lexical episode with the highest significance. Finally, for each Lexical Episode, we test iteratively whether deleting the first or last element of a lexical episode increases the significance. If this is the case, the lexical episode is pruned.

4. Visualization

To allow an exploratory analysis of the Lexical Episodes, we introduce a novel visualization that follows the design principle of overview first, zoom and filter, detail on demand [Shn96]. We first describe the layout algorithm and, second, the possibilities of user interaction.

To demonstrate the capabilities of our algorithm and visualization, we use the third Presidential debate between President Barack Obama and Governor Mitt Romney on October 22, 2012. Our approach is, however, generic, i.e. it is independent of the type of textual data.

4.1. Layout of the Visualization

Figure 2 demonstrates the visualization of the Lexical Episodes. In general, we decide to layout the textual data (in this case the presidential debate) vertically as a sequence of sentences within paragraphs (in this case within utterances). The horizontal layout dimension is used to place the sequence of words within the sentences. Lexical Episodes are visualized as a vertical bar to the left of the textual data. Each Lexical Episode is denoted an Episode Bar. These Episode Bars span from the first to the last occurrence of a relevant keyword in a sentence within the defined lexical episode. All further occurrences within the episode are visualized with a black horizontal line. The episodes for different unigrams or ngrams can overlap in word index positions. Hence, the visual representations of episodes have to be arranged using the open space to the left of the text in a way that overplotting is avoided. The basic principle we apply here is to make use of the level of significance: Episode Bars with higher significance will be ordered first; episode bars with lower significance will be arranged subsequently to the left.

In a subsequent step, labels for each episode are added. For each Lexical Episode, one label containing the lemma is added to the left of all Episode Bars. In addition, the label is connected with a visual link to the episode. The position is dependent on the average position of the keyword within the lexical episode. This is demonstrated in the Panel B in Figure 2.

4.2. User Interaction

The visualization has been implemented in Java making use of the piccolo2D library [BGM04] for user interaction. We offer different possibilities to perform generic exploration tasks.

Filtering: Analysts can change the threshold of significance to change the number of lexical bars to be displayed. This is illustrated in the top panel in Figure 2. With each change in significance, both the episode bar positioning as well as the size of labels is adjusted automatically.

Zooming: The zooming functionality goes beyond a mere geometrical zoom. When zooming in, the gray boxes representing the text increase their size at a linear level. The episode bars also increase their size linearly along the y-axis, but grow in sublinear manner along the x-axis. Consequently, the more detailed the view, the smaller the episode bars in comparison to the text. Figure 3 as well as the Panels A, C, and D in Figure 2 demonstrate this principle. Additionally, the textual data is subject to a semantic zoom. When

zooming in, first light gray horizontal bars appear within the dark gray boxes. Each such bar represents one sentence. Then, from a further zoom level on, the text will also appear. **Highlighting:** Clicking on an episode bar, all word occurrences belonging to this episode will be highlighted. For example, see the word "China" in Panel C in Figure 2.



Figure 3: Different zoom levels of the same text snippet. The breadth of the episode bar decreases in relation to the text representation on the right.

5. Use Cases

During the development of the visualization we conducted usability tests and interviews with several researchers from political science. This helped us to gain a better understanding of the real user needs, discover potential issues, and iteratively improve the design. Figure 2 demonstrates several analysis use cases we could identify as relevant for political scientists. Panel A shows that certain categories, such as in this case sentiments, can be highlighted in the zoomed-out version to make interesting patterns appear. Another use case is the detection of anomalies in the discourse structure. Panel D shows a section within the tv debate where the turn taking is characterized by very short utterances and sentences. As also indicated by the label, it is a section of crosstalk, which could be identified easily in the distant view. Panel B demonstrates how the most significant episodes cluster into thematic sections. The last thematic section is about the economical competition and collaboration with China. This sudden topic shift can be identified from the episode structure in the zoomed-out view, see Panel C. In the detailed view it becomes clear that this shift was explicitly triggered by the moderator.

6. Conclusion and Future Work

In this paper, we have introduced a novel automated and visual analytics approach for exploratory text analysis. In particular, we propose a method to combine "distant reading" with "close reading". For instance, the Lexical Episode Plots can be used to derive hypotheses based on the content and structure of the textual data. Example use cases from political science demonstrate the usefulness of the method. Additionally, since the Lexical Episode Plots are independent from language and type of textual data, it can be applied by any scholar of any discipline for exploratory data analysis.

Acknowledgments

This work was funded by the German Federal Ministry of Education and Research under grant 01461246 "VisArgue" and the research project "Visual Analytics of Text Data in Business Applications" at the University of Konstanz.

References

- [AGCH11] ARCHAMBAULT D., GREENE D., CUNNINGHAM P., HURLEY N.: ThemeCrowds: Multiresolution summaries of twitter usage. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents* (2011), ACM, pp. 77–84. 2
- [BGM04] BEDERSON B. B., GROSJEAN J., MEYER J.: Toolkit design for interactive structured graphics. *IEEE Trans. Software Eng.* 30, 8 (2004), 535–546. 3
- [CGW*09] CHANG J., GERRISH S., WANG C., BOYD-GRABER J. L., BLEI D. M.: Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (2009), pp. 288–296. 2
- [DWCR11] DOU W., WANG X., CHANG R., RIBARSKY W.: ParallelTopics: A Probabilistic Approach to Exploring Document Collections. In *2011 IEEE Conference on Visual Analytics Science and Technology, VAST 2011, Providence, Rhode Island, USA, October 23-28, 2011* (2011), IEEE, pp. 231–240. 2
- [DWS*12] DOU W., WANG X., SKAU D., RIBARSKY W., ZHOU M.: LeadLine: Interactive Visual Analysis of Text Data through Event Identification and Exploration. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2012, Seattle, Washington, USA, 14-19 October 2012, part of VisWeek 2012* (2012), IEEE, pp. 93–102. 2
- [GMFLJ03] GALLEY M., MCKEOWN K., FOSLER-LUSSIER E., JING H.: Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 7-12 July 2003, Sapporo Convention Center, Sapporo, Japan* (2003), Hinrichs E. W., Roth D., (Eds.), ACL, pp. 562–569. 1
- [HCJ14] HOQUE E., CARENINI G., JOTY S.: Interactive exploration of asynchronous conversations: Applying a user-centered approach to design a visual text analytic system. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (Baltimore, MD, 2014), Association for Computational Linguistics, pp. 45–52. 2
- [Hea94] HEARST M.: Text tiling: A quantitative approach to discourse segmentation. In *Meeting of the Association for Computational Linguistics* (1994). 1
- [Hea97] HEARST M. A.: Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23, 1 (1997), 33–64. 1
- [KJWE14] KOCH S., JOHN M., WÖRNER M., ERTL T.: VarifocalReader – In-Depth Visual Analysis of Large Text Documents. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 12 (2014), 1723–1732. 2
- [LZP*09] LIU S., ZHOU M. X., PAN S., QIAN W., CAI W., LIAN X.: Interactive, Topic-based Visual Text Summarization and Analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (New York, NY, USA, 2009), CIKM '09, ACM, pp. 543–552. 2
- [MH91] MORRIS J., HIRST G.: Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics* 17, 1 (1991), 21–48. 1
- [MH13] MURALIDHARAN A., HEARST M. A.: Supporting exploratory text analysis in literature study. *Literary and Linguistic Computing* 28, 2 (2013), 283–295. 2
- [O’C14] O’CONNOR B.: Mitextexplorer: Linked brushing and mutual information for exploratory text data analysis. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (Baltimore, MD, 2014), Association for Computational Linguistics, pp. 1–13. 2
- [RHD*12] ROHRDANTZ C., HAO M. C., DAYAL U., HAUG L.-E., KEIM D. A.: Feature-based visual sentiment analysis of text document streams. *ACM TIST* 3, 2 (2012), 26. 1
- [Sch12] SCHMIDT B. M.: Words Alone: Dismantling Topic Models in the Humanities. *Journal of Digital Humanities* 2, 1 (2012). 2
- [Shn96] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages* (Washington, WA, 1996), IEEE Computer Society Press, pp. 336–343. 2