

Stylistic Text Segmentation

Paul J. Chase
Illinois Institute of Technology
3300 South Federal Street
Chicago, IL 60616-3793, US
chaspau@iit.edu

Shlomo Argamon
Illinois Institute of Technology
3300 South Federal Street
Chicago, IL 60616-3793, US
argamon@iit.edu

ABSTRACT

This paper focuses on a method for the stylistic segmentation of text documents. Our technique involves mapping the change in a feature throughout a text. We use the linguistic features of conjunction and modality, through taxonomies from Systemic Functional Linguistics. This segmentation has applications in automated summarization, particularly of large documents.

Categories and Subject Descriptors

I.7 [Document and text processing]: Segmentation

General Terms

Algorithms, Languages

Keywords

Systemic Functional Linguistics, Text Segmentation

1. INTRODUCTION

This paper introduces a method for automatic stylistic segmentation of a document. The intention is to determine introductory, conclusive or transitional segments within a document; the motivation being that such sections should be valuable in text summarization tasks. Current methods of text segmentation [8, 2, 5] focus on finding topical shifts between segments; these methods have excellent utility for document indexing and searching, however they fall short when applied to the task of document summarization [7]. One problem stems from the typical summarization technique of extracting salient sentences [3] from each section; in a lengthy document, this yields either an unacceptably long summary or one which fails to represent possibly important topics [4]. A second problem arises in that both introductory and conclusive paragraphs have great topical similarity with their preceding or following passages, thus often represented as a single section by traditional methods.

We show the results of stylistic segmentation on formal scientific writing, analyzing their agreement with the true headings in each document. Our technique involves mapping the changes in types of conjunction and modality expressed over the course of a text; this differs from previous approaches of finding new terms [2] or lexical cohesion [5] methodologies. For this purpose we utilize Systemic Functional Linguistic features as a theoretical framework; however, the method may conceivably be used with any similar topic-independent feature set.

2. APPROACH

What is lacking from topic-based segmentation may be replaced with stylistic segmentation; for this we use features independent of the topic of sections. These features are derived from the taxonomies of Systemic Functional Linguistics (SFL) [1]. The first feature group used is conjunction, mapping the way clauses or sentences are linked to each other. Second is the expression of modality, concerning modal verbs like can, must, and should, used to modify verbal events. Contained within modality lies the third feature group exploring the degree of modality; e.g. ‘can’ is weaker than ‘must’.

Systemic Functional Linguistics represents language as a message, in the form of independent *systems* describing various stylistic choices. Each system contains *options* representing the author’s chosen stylistic representation; the system of EXPANSION, for instance, describes conjunction. Within this system exist three options: Extension, which links clauses giving different information together; Enhancement, which qualifies information in one clause by another and Elaboration, which deepens a clause by clarification or exemplification.

The two subsystems of modality are TYPE and VALUE. The former represents a choice between Modalization, expressing the likelihood or frequency of an event, and Modulation, expressing its ability or necessity; VALUE has the three options High, Median and Low and accordingly represents the degrees of modality.

Our first step was to extract the section headings from the pdf documents by using links within the documents themselves. The section headings are then replaced with an inert¹ identifier to ensure they are not inadvertently affecting the automatic segment detection. The documents were then an-

¹Inert – in this context – indicates that the identifier is not lexical in nature and cannot be recognized by the section detection software. It is, however, easily found when comparing found to true sections.

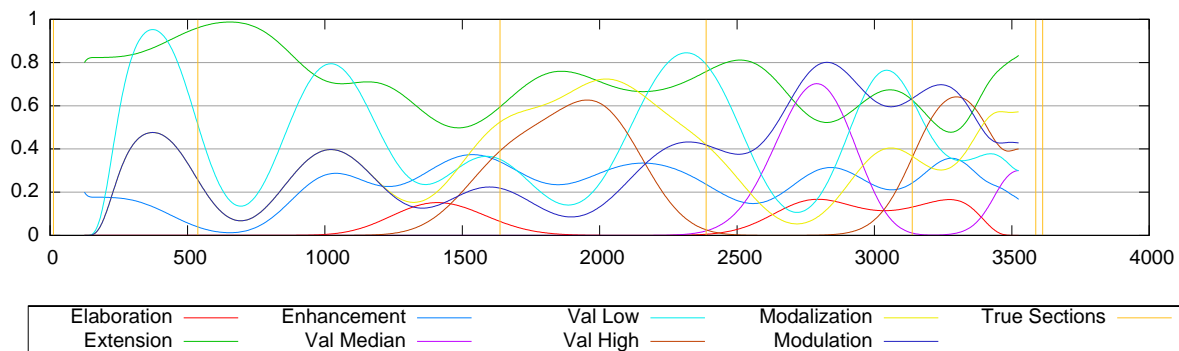


Figure 1: An example article. Vertical lines indicate true section breaks.

alyzed according to 250-token windows, sliding the window one 25 token section at a time (90% overlap). Within each of these windows the relative frequencies of each option within a system were recorded:

$$RF_i = \text{count}(\text{Option}) / \text{count}(\text{System})$$

This generates a vector for each option, from which its slope is estimated. If the rate of change within a window is high enough then each of the 10 25-token sections within it is marked as a likely section break by that feature. Our system relies on two thresholds: $tSlope$ determines what is considered a high slope while $tAgree$ is what fraction of features must have at least this slope to indicate a section boundary.

3. EVALUATION

Document segmentation is a highly subjective task which usually involves finding the boundaries by hand, a laborious process. To avoid such, we utilize highly structured text in our evaluation; our corpus consists of 368 formal scientific articles from four different fields. Each article is divided based on the authors' placement of section headings.

The first step in this investigation was to evaluate the viability of the system manually; to that effect graphs were constructed of the features changing over time - see Figure 1. The vertical lines are the true sections; they tend to correspond to regions where many of the feature curves have a high slope. Further evaluation was performed using the familiar F-measure, with balanced precision and recall. This measure is somewhat crude for this task (see [6]), however we believe it remains useful for preliminary study.

4. RESULTS

Figure 1 is a visual indication of the generated data; vertical lines indicate true sections, while the curves represent the value of each feature used. The thresholds were used from 0 to 2 for $tSlope$ and 0 to 1 for $tAgree$ to obtain the F-measures mapped in Figure 2. Our best result, $F = 0.635$, occurs with $tSlope = 1.6$ and $tAgree = .3\%$; here precision and recall are $P = 63\%$ and $R = 65\%$. The precision indicates many false sections are found; we hypothesize that this is due to stylistic changes independent of marked section boundaries. The recall is expected in that this system is not intended to find topical boundaries, which are quite prevalent in some documents. Previously reported results on the textTiling [2] algorithm were $F = .3556$ on a single

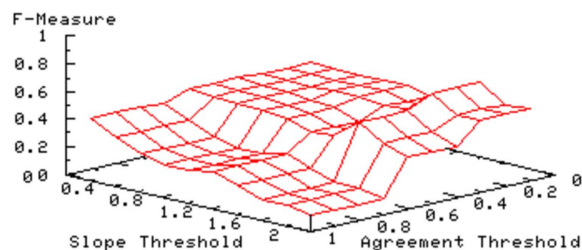


Figure 2: Plot showing F-measures.

document; the vecTile [5] improves this to $F = .515$, on a corpus of 5 popular-science papers.

5. CONCLUSIONS

We have presented a new technique for linguistic analysis: mapping the change of linguistic features over the course of a document. To explore the power of this technique it was applied with Systemic Functional Linguistic features to a subtask of document segmentation. Even as a generic segmentation tool, this technique shows merit; further study will examine its particular strengths and weaknesses.

6. REFERENCES

- [1] M. A. K. Halliday. *An Introduction to Functional Grammar*. Longman, London, 1994.
- [2] M. A. Hearst. Textiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.
- [3] V. M. J. Goldstein, M. Kantrowitz and J. Carbonell. Summarizing text documents: sentence selection and evaluation metrics. pages 121–128. SIGIR, 1999.
- [4] M. Y. Kan, J. L. Klavans, and K. R. McKeown. Linear segmentation and segment significance. *6th International Workshop of Very Large Corpora*, pages 197–205, August 1998.
- [5] S. Kaufmann. Cohesion and collocation: using context vectors in text segmentation. *37th Annual Meeting of the Association of for Computational Linguistics (Student Session)*, pages 591–595, June 1999.
- [6] L. Pevzner and M. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, 2002.
- [7] R. D. B. R. Angheluta and M. F. Moens. The use of topic segmentation for automatic summarization. pages 66–70, Philadelphia, Pennsylvania, USA, July 2002. Workshop on Automatic Summarization.
- [8] H. G. Silber and K. F. McCoy. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics*, 28(4):487–496, 2002.