Statistical Natural Language Processing Method for Variant Texts Segmentation

Maki Miyake (mmiyake@dp.hum.titech.ac.jp)
Hiroyuki Akama (akama@dp.hum.titech.ac.jp)
Masanori Nakagawa (nakagawa@nm.hum.titech.ac.jp)

Department of Human System Science, Tokyo Institute of Technology, 2-12-1, Ookayama, Meguro-ku, Tokyo, Japan

Introduction

It is well known that some techniques have already been developed to automatically subdivide texts into multi-paragraph subtopic passages, such as TextTiling methodology proposed by Hearst. However, an additional algorithm is needed to perform a similar task for parallel or variant texts, because ambiguous and complicated traces of cross citation among them might often generate some sinuous patterns of lexical co-occurrence that make fuzzy the boundaries of units of coherent episode. In other words, we are confronted with a sort of Frame question of how we partition off the texts to respect their own genealogy and avoid irrelevant interpretation of source reference.

In this paper, we propose a new statistical natural language processing method to partition off the variant texts. The Parallel Synoptic Tables (PST) in the Synoptic Gospels, Matthew, Mark and Luke are taken as examples of variant texts to which our new method will be applied. The method makes it possible for us to obtain the Computed Synoptic Tables (CST) by providing us with new objective segmentations of the parallel texts in Synoptic Gospels.

The Proposed Method

The method we present is called SynopticPatch, which consists of the combination of 1) N-gram calculation, 2) Windowing data gathering and 3) Application of TextTiling method. The aim of this method is to produce an objective collateral segmentation of the parallel or variant texts by using computational criteria.

All the N-gram instances common to the variant texts are gathered so as to make an exhaustive list of the pair strings of words that are correspondent to each other at their parallel positions. We set up a set of synchronized extending windows for each parallel N-gram instance to be centered in. In the operation rule, each window is supposed to stop the extension in its size if the border meets some critical positions such as previous or next N-gram instances, punctuation marks while recording the frequency data of the co-occurring words one by one and simultaneously in each of the variant texts. We fix the segmentation point by the threshold which is referred to the correlation coefficient between the word frequency vectors generated from each corresponding window instance.

The Computed Synoptic Tables

The first step of SynopticPatch is to calculate the overall N-gram data under the condition of ($N \geqq 3$). The words overlapping between the texts are classified by the four combination patterns (A,B,C,D in Figure1). Secondly, a set of synchronized windows, changing its size depending on each parallel n-gram instance, records the frequencies of the co-occurring words one by one and at the same time until it stops extending its size with meeting of the previous or the next pericope. Finally, we calculate at every step of the window extension the correlation coefficient. We compute at each word position the mean of the correlation coefficients obtained from all the series of scores with plains at their centers. In the case of the Synoptic Gospels, we empirically set the threshold as 0.5 for the cohesion score graph of the Computed Synoptic Tables (CST).

As the result, the index of difference between the Parallel Synoptic Tables (PST) and the CST can be defined by the distribution of the words into the 7 categories. The effects of the new combinations are clearly revealed by the diminution in quantity of some textual overlaps.

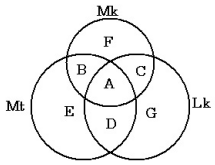

Figure 1: 7 categories

Conclusion

The SynopticPatch enables us to automatically partition off the parallel and variant texts. For the Synoptic Gospels, our method made it possible to generate the new synoptic tables based upon the computed segmentation of biblical episodes. Experimental results indicate that the method will lead us to propose a new interpretation of Synoptic Gospels at least in the framework of computational humanities.

Acknowledgments

The work is performed as a part of the 21st Century Center of Excellence Program "Framework for Systematization and Application of Large-scale Knowledge Resources".

References

Minsky, M.L., (1975), A Framework for representing knowledge, The Psychology of computer vision, pp.211-277.
Hearst, Marti A., (1997), TextTiling: “Segmenting text into multi-paragraph subtopic passages”, *Computational Linguistics* 23, pp.33-64.
Conzelmann, H. & Lindemann, A.. Interpreting The New Testament, trans. by Siegfried S. Schatzmann, Hendrickson Publishers,45-53, 1988.