# 論文の内容の要旨

## 論文題目： A Study on Unsupervised Segmentation of Text Using Contextual Complexity（文脈の複雑性に基づく教師なし文分割に関する研究）

## 氏 名： 靳 志輝

## 要旨

Language is essentially non-random; there are hidden regularities that can be captured by exploring the distributional patterns existed in the language data. Considerable research in language acquisition has addressed the extent to which basic aspects of linguistic structure might be discovered on the basis of statistical property of language. From engineering point of view, many applications in natural language processing have shown that the computers equipped with some knowledge of human language can be helpful tools. The innate statistical properties of the language make it possible to automatically fetch the useful knowledge from huge amount of language data which have become available nowadays.

My research on unsupervised text segmentation focuses on the structure of the language with regards to morpheme and word. Linguistic structuralist Zellig S. Harris claimed the regularity that English morpheme/word boundaries can be detected from changes in the complexity of phoneme sequences. In my research, I verified Harris's hypothesis in a fundamental manner and addressed the question of why and to what extent the hypothesis holds in different languages. The hypothesis was reformulated from an information theoretic viewpoint, and then several large-scale experiments were conducted to evaluate it in different languages. The result I got was of interest both from engineering viewpoint and linguistic viewpoint.

The hypothesis was firstly applied to written Chinese text to do unsupervised word segmentation. Since languages change over time,

detecting and handling unknown words properly become a crucial issue in today's practical natural language processing. Thus, the research to do unsupervised Chinese word segmentation without using any pre-segmented data but based on raw text, will shed light on how can we deal with unknown words. In my research, a large-scale experiment was conducted and the result I got was quite comparable to other unsupervised segmentation methods.

Secondly I test the hypothesis in phoneme sequence of spoken English and Chinese. I found that the hypothesis holds well for morphemes in both English and Chinese. However, I obtained contrary results for English and Chinese with regard to word boundaries; this reflects a difference in the nature of the two languages.

Furthermore, the limitations of the Harris's method are found in error analysis of segmentation result. In order to improve the segmentation performance, several models to do global optimization on the segmentation result was proposed. An experiment conducted in written Chinese text significantly improved the performance and arrived in the state-of-the-art result in unsupervised Chinese word segmentation.