International Journal of Technology and Human Interaction
Volume 15 • Issue 1 • January-March 2019

Information Structure Parsing for Chinese Legal Texts:
A Discourse Analysis Perspective

Bo Sun, Hefei Normal University, Hefei, China

ABSTRACT

Information processing is one of the main concerns in the field of artificial intelligence, because it can benefit many related downstream tasks. To facilitate information processing, information structure parsing is assumed to be of great significance. This article proposes a discourse analysis based approach so that information structure of Chinese legal texts can be recognized automatically. This article employs Discourse Information Theory to explore information features of Chinese legal texts. The texts used in this study include 6 types, each type containing 60 training texts and 30 testing texts. After that, a set of rules is formulated to classify legal texts and identify the categories of information units. Finally, to examine the performance of the rules, a comparison is made by designing a Support Vector Machine classifier and a Viterbi algorithm decoder. The experiment demonstrates that the rule based approach outperforms the statistics based approaches. This research suggests that discourse analysis may provide some linguistic features conducive to discourse parsing.

KEYWORDS

Discourse Analysis, Discourse Information Theory, Information Recognition, Processing Rules, Text Classification

INTRODUCTION

With the advancements of the *Free Access to Law Movement* across the world (Greenleaf, 2011), legal information becomes more and more easily available for lay people, legal professionals and law makers. Portals and databases, including but not limited to *China Judgements Online* (http://wenshu.court.gov.cn/), *Beida Fabao* (http://www.pkulaw.cn/) and *Westlaw*, can provide users with thousands of judicial cases, court decisions and regulations.

Meanwhile, text-based legal information processing has always been a heated topic within the field of artificial intelligence and law (Bench-Capon et al. 2012). Researchers pay much attention to such issues as legal information retrieval (Rissland & Daniels, 1995; van Opijnen & Santos, 2017), legal information extraction (Moens et al., 1999; Webber et al., 2005), Online Dispute Resolution (Carneiro et al., 2014; Zeleznikow, 2017) and eMediation (Jelali et al., 2015). From legal texts, litigants can find out possible outcomes of an impending trial and evaluate their potential risks; attorneys can foretell what counterclaims will probably be put forward by opposing counsels; judges can have a good understanding of other judges' inferences and attitudes in similar cases, especially when they are faced with high-profile, complicated and difficult cases.

However, as the exponential increase of legal texts in the Internet age incurs an overload of information (Koniaris et al., 2017; Opijnen & Santos, 2017), useful information has to be extracted through Text Mining technique (Andrade & Santos, 2017). Although text classification can help to

DOI: 10.4018/IJTHI.2019010104

exclude legal texts irrelevant to users' queries (Ashley & Brueninghaus, 2009), redundant information is still included in the remaining texts, as users might just need some part of a text. Therefore, to enhance user satisfaction, legal texts should be processed in a more fine-grained way.

To this end, techniques from several disciplines like linguistics, machine learning and Natural Language Processing (NLP) can be exploited. Among all these, discourse analysis may offer some special insight, because discourse is not a random combination of sentences or words. Moens et al. (1999) made rules through discourse analysis and built the SALOMON system to abstract Belgian criminal cases. Kunc et al. (2013) believed that knowledge of discourse structure in dialogue could help to optimize human-computer interaction. Taboada et al. (2009) demonstrated that the accuracy of sentiment analysis could be boosted if discourse structure was taken into account. Lin et al. (2005) even used discourse-based features such as cue phrases to segment lecture videos which are a kind of multimodal discourse. All these studies manifest the usefulness of discursive factors in text processing.

In this paper, the author addresses the problem of identifying information structure of Chinese legal texts automatically by proposing a discourse analysis based approach. The analysis is made within the framework of Discourse Information Theory (Du, 2014).

The remainder of this article is structured as follows. Section 2 introduces the background of this study, including a brief overview of information theories on language, theoretical research on discourse parsing and the argument over rationalism and empiricism in NLP. Section 3 provides the theoretical framework of this study and articulates the proposed approach. Section 4 describes the methodology of this research. Section 5 applies Discourse Information Theory to the analysis of Chinese legal texts and presents their information characteristics. Section 6 formulates information rules based on these characteristics. Section 7 examines the efficacy of the rules with two experiments. Section 8 is the concluding part.

BACKGROUND

A Brief Overview of Information Theories on Language

Undoubtedly, traditional research on language information is rooted in Shannon's Information Theory (Shannon, 2001), which has laid a solid foundation for modern information technology. However, Morin (2013) pointed out that the theory failed to show adequate concern for social science, because information also bore social characteristics. In addition, Yue (1996) argued that the theory described information only in a quantitative manner without considering the meaning of information.

Linguistic research on information is mostly guided by functionalism. A renowned taxonomy of language information is given information and new information (Halliday & Matthiessen, 2004). The former is what the addressor believes is known to the addressee, while the latter is the information that the addressor assumes is not known to the addressee. This dichotomy has been accepted and developed by many linguists. For example, Prince (1981) classified given information into the inferable and the evoked. Brown and Yule (2003) divided new information into the brand-new and the unused and grouped evoked information into the situational and the textual. Although such functionalist approach can provide at many as 6 types of information, it still seems inadequate to portray the various kinds of information in legal texts. Besides, a person without proper linguistic training can hardly tell the differences among these categories.

The "5W" Model in communication research generalized information with five interrogatives, namely "who", "says what", "to whom", "in what channel", and "with what effect" (Lasswell, 1948). The categorization was later developed to "7W" Model by adding "under what circumstances" and "for what purpose" (Braddock, 1958). This classification of information conforms to human beings' intuitive understanding of information, but it delineates only the linear structure of language information.