The International Journal of Speech, Language and the Law

IJSLL (PRINT) ISSN 1748-8885
IJSLL (ONLINE) ISSN 1748-8893

PhD Abstract

A study on the automatic representation mechanism of legal discourse information

Hong Wang

School of Foreign Languages,
Shandong Normal University,
No. 88, Wenhuadong Road,
Lixia District,
Ji'nan 250014
China

Awarding Institution:	Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, China
Date of Award:	27 December 2019

KEYWORDS: DISCOURSE INFORMATION, AUTOMATIC ANNOTATION, LEGAL REASONING, PATTERNS OF LEGAL DISCOURSE INFORMATION, MULTI-PROCESS HANDLING MECHANISM

Contact
email: wanghonghelen@126.com

IJSLL VOL 28.1 2021 133–138 https://doi.org/10.1558/ijsll.20601

With the implementation of judicial modernisation, AI-assisted judicial practice has become a trend in China for being able to enhance efficiency. The deep integration of AI in judicial practice largely relies on the development of automatic processing of legal language, which can be promoted by annotated legal corpora. This study advances the automatic processing of legal language at the discourse level.

As a concern of research in natural language processing and computational linguistics, annotation gets attention for increasing the range of linguistic phenomena in a corpus and for helping to realise information extraction from a corpus more accurately. Since manual corpus annotation is time-consuming and laborious, some researchers have turned to focusing on the realisation of automatic annotation and have already achieved the goal at lexical, syntactic and semantic levels. Automatic annotation at the discourse level has not been fully realised due to a lack of applicable linguistic representations, hence its application in some intelligent systems, including intelligent court systems and intelligent medical systems, has been postponed.

In the construction of a Chinese intelligent court system, a large-scale legal corpus is a preferred source of linguistic knowledge, if data annotated at the discourse level is available, for the reason that discursively annotated data may assist courtroom trials by providing automatic case classification and summarisation, more accurate information extraction, assisting prediction of case decisions and so on.

Legal communication is a dynamic and inferential process that is reflected in various language patterns. Discourse information underlies linguistic forms and is exhibited by different linguistic forms, which influences annotation at the discourse level. Thus, it is necessary to seek out and annotate discourse information in legal communication if AI language processing is to be part of judicial practice.

Legal communication interweaves with legal reasoning and uses rules to 'draw conclusions about the existence of particular rights or duties in a given situation' (Vandevelde 2011: 55). Reasoning and information are synchronous and inseparable in linguistic interactions (Benthem 2008), including communication in legal contexts. That is to say, information is embedded into language in communication, and, as the logical basis of information, reasoning develops in parallel with information and is reified in the compilation of information. Reasoning is an indispensable segment in acquiring information (Tang 2003), so clarifying discourse information based on reasoning is the prerequisite to automatic annotation at discourse level.

Therefore, representing discourse information in legal communication on account of legal reasoning has become the focus of the present study. The research objective is that the representation of discourse information will help in realising

automatic annotation at the discourse level to promote legal corpus annotation, and hence contribute to automatic processing of legal documents by integrating discourse theory into AI-assisted judicial practice.

Three specific questions are raised to achieve the research objective, as follows:

1. What are the reasoning-oriented schemata developed in Chinese legal communication?
2. What is the discourse information template that evolves from legal reasoning?
3. How is discourse information represented automatically for annotation at the discourse level?

To accomplish the research objective, an analytical framework was constructed for the representation of discourse information in Chinese legal communication by incorporating Discourse Information Theory (DIT) (Du 2007, 2014), legal reasoning and the concepts of 'distributional hypothesis' and 'selectional preferences', which are each explained below.

DIT provides a theoretical basis and a feasible perspective in analysing and representing discourse information. It defines a discourse as a hierarchical structure consisting of information units each of which stands for the smallest meaningful unit. The analysis focusing on information units, information knots and information development constructs the macro-structure of discourse information, and the analysis concerning three kinds of information elements (Process, Entity and Condition) frames the micro-structure of discourse information. Information elements are the constituent components of information units.

Legal reasoning in Chinese legal discourse mainly follows three reasoning patterns: factual reasoning, regulatory reasoning and adjudicatory reasoning (Wang 2013). These three main reasoning patterns are realised through various modes of reasoning process. Factual reasoning is employed to summarise legal facts and to identify relations between facts. Regulatory reasoning can help judges to define litigants' rights and obligations, to interpret laws and regulations and to instruct judicial proceedings. Adjudicatory reasoning is used to infer a verdict on the basis of laws and regulations (the major premise) and legal facts (the minor premise).

According to the distributional hypothesis, all linguistic elements are divided into different groups whose relative occurrence can be stated exactly. That is, linguistic elements can be grouped according to their distributional behaviour if they distribute similarly. Harris (1954, 1991) argues that it is possible to describe the whole distributional structure of language with respect to the distributional behaviour of a linguistic element relative to other elements, and that such distributional accounts of linguistic phenomena are without the intrusion of other features.

However, the distribution of linguistic elements is restrained by selectional preference, which means a certain set of words being more likely to be used in combination with words of another particular set (Harris 1991: 25). Probably every word in a sentence and every sentence in a discourse has a unique selectional preference of co-occurrences and unique inequalities from most likely to least likely co-occurrents (Harris 1991: 64). Selectional rules define selectional relations between linguistic elements (Chomsky 1965: 113).

From the theorised three dimensions of discourse, i.e., the information element, information unit and information level, the present framework provides a guide for exploring the features and collocations of discourse information, the governing scope of each information level, and legal discourse information patterns, for automatic annotation.

Methodologically, a corpus-based qualitative analysis was conducted. The data used were taken from the Corpus for the Legal Information Processing System with manually annotated data as a reference and unannotated data as test material for automatic annotation. Automatic annotation was realised through Java programming in Microsoft Word formatting by following seven steps, namely: pretreatment of texts, automatic recognition of genres, automatic generation of the Kernel Proposition, automatic recognition and annotation of information elements, automatic identification and annotation of information units, automatic definition of the governing scopes of information levels and automatic annotation of information levels.

A legal document to be processed was pretreated firstly to keep the indispensable parts, including the main body, the title and the case number etc. The text genre was then recognised automatically by referring to the heading and level-1 information units, and a Kernel Proposition was generated as well. A problem influencing the remaining four steps was how to validate an information unit containing no verb or more than one verb. The problem was tackled by following the Multi-Process Handling Mechanism designed in the present thesis, as shown in Figure 1.

Relations between information units were manifested in hierarchical and parallel relations. Information units and information levels can reflect legal reasoning from different perspectives.

The major contributions of the study lie in representing Chinese legal discourse information on the basis of DIT and legal reasoning and hence realising automatic annotation of legal discourse information in a corpus of legal texts. The annotation was realised at three levels, i.e., information elements, information units and information levels, both linearly and hierarchically. The automatic representation helps in generating a Kernel Proposition to extract the main idea of a discourse and in recognising and annotating information elements, information units and

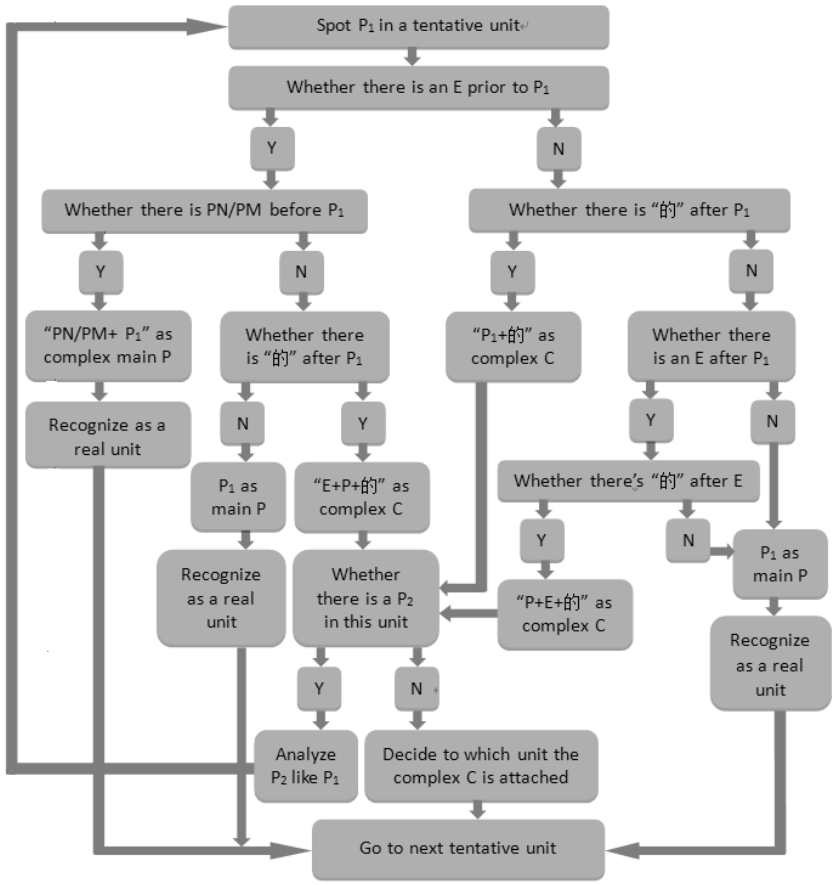

Figure 1: The Multi-Process Handling Mechanism

information levels. The types and governing scopes of information units are recognised with the help of the Multi-Process Handling Mechanism constructed in the study to deal with information units containing no verb or more than one verb, which might hinder the automatic recognition of an independent information unit and the governing scope of an information level. In addition, three new subtypes of discourse information elements – 'Modality (PM)', 'Causativity (PSV)', and 'Conjunctives (Cj)' – were put forward.

This study represents legal discourse information so that automatic annotation can be realised, which can speed up the annotation process and save labour and time. It promotes the application of DIT in corpus annotation at the discourse level and sheds light for the development of discourse structure analysis in natural language processing, especially in automatic summarisation, topic extrac-

tion and question answering. It is also hoped that the study may provide useful forensic linguistic references for the automatic annotation at the discourse level of other genres and other languages, and thus contribute to expanding the applications of forensic linguistics research.

Acknowledgments

I am very grateful to Prof. Du Jinbang, Prof. Xu Zhanghong, Prof. Yuan Chuanyou, Prof. Zhao Junfeng, Prof. Chen Jinshi, Prof. Ge Yunfeng, Dr Zhang Shaomin, Dr Guan Xin, Dr Yu Xinbing, Dr Liu Juan and Dr Sun Bo for their help and encouragement. This work was supported by the National Social Science Fund of China (16BYY064) and the Social Science Project of Shandong Province (21CYYJ14, 13CWXJ23).

References

Benthem, J. (2008) Logic and reasoning: Do the facts matter? *Studia Logica* 88(1): 67–84. https://doi.org/10.1007/s11225-008-9101-1

Chomsky, N. (1965) *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Du, J. (2007) A study of the tree information structure of legal discourse. *Modern Foreign Languages* 1: 40–50.

Du, J. (2014) *On Legal Discourse Information*. Beijing: People's Publishing House.

Harris, Z. (1954) Distributional structure. *Word* 10(2–3): 146–162. https://doi.org/10.1080/00437956.1954.11659520

Harris, Z. (1991) *A Theory of Language and Information*. Oxford: Clarendon Press.

Tang, X. (2003) *A Logical Analysis of Cognition*. Chongqing: Southwest China Normal University Press.

Vandevelde, K. J. (2011) *Thinking Like a Lawyer: An Introduction to Legal Reasoning* (2nd edn). Boulder, CO: Westview Press.

Wang, H. (2013) *Legal Logic*. Beijing: China University of Political Science and Law Press.