報告番号	※甲　　第　　号

主　論　文　の　要　旨

論文題目　Coordination Analysis and Term Correction for Statutory Sentences using Machine Learning
（機械学習による法令文の並列構造解析及び用語校正）

氏　　名　　山腰　貴大

論 文 内 容 の 要 旨

Laws constitute an essential infrastructure that sustains and improves human society. Statutes stipulate laws by natural language, and they are continuously updated as society changes. Since statutes prescribe the rights and duties of people, statutory sentences are not allowed to contain errors or inconsistencies. To keep high consistency, they are written in accordance with their specific wordings, technical terms, and sentence structures. Therefore, we need to have sufficient knowledge and experience to write and understand statutory sentences appropriately.

In the case of Japan, a number of writing rules and customs have been established since the Meiji era through legislation (works of interpreting, producing, and updating statutes). Furthermore, the Japanese government has legislation bureaus that strictly examine draft bills whether they are written in concordance with the legislation rules. For these two reasons, it is quite important to follow the legislation rules when we handle Japanese statutory sentences, which may burden legislation officers with comprehensive and strict writing. Another point to be noted in Japanese statutory sentences is that they tend to be quite long and complex. One big factor for this characteristic is coordinate structures in Japanese statutory sentences. A coordinate structure is a sentence structure that enumerates multiple things in parallel. In Japanese statutory sentences, such coordination often appears with hierarchy, that is, a coordinate structure contains other coordinate structures inside. The legislation rules stipulate the hierarchy of coordinate structures; in other words, we need to mind the rules when we understand the coordination of statutory sentences. Overall, we identify two subjects that will be obsta-

cles in handling Japanese statutory sentences: strict compliance with legislation rules and complex hierarchical coordinate structures.

In this thesis, we study two themes to provide solutions for the two subjects: coordination analysis and legal term correction. Coordination analysis identifies scopes of conjuncts (phrases in parallel) in a given sentence. With this information, we can simplify long and complex statutory sentences, which supports any person and system that has trouble understanding such statutory sentences. Thus, we place this study on a quite fundamental one for further sentence processes. The second theme, legal term correction, is located to a practical study that aims at drafting statutory sentences. The legislation rules define distinct usage of certain similar legal terms, which should be fulfilled in drafting statutory sentences. Our legal term correction finds misused legal terms and offers correction ideas for them, that is, this is proofreading specialized in legal terms with distinct usage.

The approaches in this thesis are a combination of deterministic legislation rules and machine learning technologies. It is reasonable to import the Japanese legislation rules to the approaches as deterministic rules because these rules are well-established and strictly operated by the government. We then delegate decisions based on context to machine learning methods. Both the formation of coordinate structures and the use of legal terms depend on the context around them. Since the number of context patterns is enormous to cope with deterministic rules, we rely on machine learning methods that automatically learn contexts from training data.

This thesis consists of seven chapters. Chapter 1 is the introduction of this thesis, which begins with an explanation of the legislation and Japanese statutory sentences. After identifying our studies that solve issues in handling Japanese statutory sentences, we position them among their related studies.

In Chapter 2, we describe the knowledge and techniques that are the basis of our proposed methods. First, we review the Japanese legislation rules, and then we dig into coordination and legal terms that are the subjects in this thesis. Next, we look at language models and classifiers that are the core machine learning technologies in the approaches.

In Chapter 3, we describe the study for coordination analysis for Japanese statutory sentences. We first review the background of coordination analysis including issues in current situations. We then propose a coordination analysis method for Japanese statutory sentences by comparing an existing method for them. Our method deterministically identifies the hierarchy of coordination based on the Japanese legislation rules for hierarchical coordinate structures. On the other hand, it identifies the scopes of conjuncts that compose a coordinate structure by utilizing neural language models. Here, we introduce two assumptions on coordination

that ensure the validity of conjunct scope candidates. The first assumption is the conjunct similarity, that is, two paired conjuncts have similar context. The second assumption is the conjunct interchangeability, that is, a sentence is still fluent even if we swap two paired conjuncts in its coordinate structure. We calculate scores of these two assumptions by neural language models that are aware of the context of the whole sentence. In addition, the models are trained with sequences of tokenized statutory sentences; In other words, we do not use coordination information for training. This enables us to realize a neural-based coordination analysis method for Japanese statutory sentences with limited training resources.

In Chapter 4, we describe the study for legal term correction for Japanese statutory sentences. As same as Chapter 3, we first review the background and needs of legal term correction. Since the legal term correction task has not been studied yet to the best of our knowledge, we first define this task, and then we consider its characteristics. Next, we propose two approaches for the legal term correction task. The first approach uses Random Forest classifiers, which assigns a trained Random Forest classifier to each legal term set. Here, each classifier is optimized by its corresponding legal term set, and thus high prediction performance is expected. Furthermore, we earn knowledge on legal term correction from optimized parameters and feature importances calculated in the training. The second approach uses a BERT classifier, where we aim to achieve further good prediction performance by utilizing the wider context capability from the self-attention mechanism and the enormous knowledge earned by pre-training. Here, we introduce a problem of two-level infrequency in the legal term correction task and a solution for it.

In Chapter 5, we attempt to apply the legal term correction methodology established in Chapter 4 to foreign statutes, namely Thai statutes. It is a global issue that statutory sentences should be written appropriately. Here, Thai legislation has rules on the usage of similar legal terms, which is the same as Japan. On the other hand, usage of Thai legal terms tends to be bound by outside-sentence contexts such as genre and year. Also, Thai legal terms sometimes appear with few adjacent words, which we do not normally observe in Japanese legal terms. Therefore, we apply additional features for Thai legal term correction to the Random Forest approach of the previous chapter.

In Chapter 6, we discuss the relationship between the studies and real-world data circulation from the viewpoints of the existence of data circulation in the studies and contributions that the studies bring.

In the final chapter, we summarize this thesis. We first organize discussions in the previous chapters and then we discuss future work and prospect of the studies.