Towards a Conceptual Search for Vietnamese Legal Text

Thinh Bui, Son Nguyen, Quoc Ho

Towards a Conceptual Search for Vietnamese Legal Text

Thinh D. Bui, Son T. Nguyen, and Quoc B. Ho

Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam
bdthinh@fit.hcmus.edu.vn, ntson@fit.hcmus.edu.vn, hbquoc@fit.hcmus.edu.vn

Abstract. In this paper, a system of search engine is built based on documents of Vietnamese legal system. The key factor of this system is the ability of indexing documents in several aspects: not only on the whole text but also on logical structures and ontologies of Vietnamese legal documents. There are two important phases in the system; firstly, focusing on the recognition of the structure of Vietnamese legal text and the extraction of ontology of documents stored in the database of Vietnam Ministry of Justice; secondly, building an automated information retrieval system for the advanced search on demand of Vietnamese citizens. We study Vietnamese legal domain in both linguistic features and patterns of recognition. Experimental result on recognizing logical structures got of 64.37% on assumption annotation, 64.15% on provision annotation and 75.76% on sanction annotation in the $F_{\beta=1}$ score on the pre-built corpus of Vietnamese Enterprise Law articles. We also evaluated the precision of classifying ontology concepts and got 86.4 % in result on testing sample set of Vietnamese legal ontology. The goal is that a search engine is proposed with indexed logical and semantic properties based on works of the first phase.

Keywords: legal text processing; legal text mining; legal ontology; information retrieval.

1 Introduction

Legal texts are a kind of document with specific characteristics, which is different from other daily-used documents due to their length and complexity [1][2]. It is very important to be able to structure and extract information automatically from legal texts containing legal articles and cases in order to meet two clearly predefined goals [3]; firstly, supporting experts to establish complete and consistent laws and secondly deploying information systems which help citizens understand laws more effectively.

In syntactic aspect, our study of Vietnamese legal texts leads to a judgment that the law sentences and paragraphs have some specific structures [1][2]. In most cases, an article which is compound by one or many sentences can regularly be divided into two of three parts: the assumption part, the provision part, and the sanction part [4]. Recognizing the logical structures of legal articles is an

important task to understand the meaning of legal domain and a preliminary step in problems such as translating a legal article into the logical form, legal text summarization, and question-answering in this domain [3]. In a legal article, the assumption part usually describes the subjects and cases in which the law can be applied, and the provision part tells mandatory regulations related to the subjects. The sanction part describes a forceful penalties in cases the subjects contravene the provision which is applied to them [4]. Therefore, the outputs of recognition of logical structures of legal text is very helpful to citizens who want to understand the law: what a legal article says, or which cases the law can be applied, or which subjects related to provision or sanction described in the legal article [1][2].

Moreover, since legal domain is highly entangled with common sense views on the nature of social events, roles, and actions, the task of building legal ontologies is essential to cover the understanding of these concepts and to help legal documents be machine readable. We presently aim to build a fulfilled Enterprise Law ontology on a schema as a subtree of LKIF core ontology schema [5][6].

In this paper, we present a conceptual search for Vietnamese legal text which use logical and semantic annotations. We also implement a system based on an open-source search engine and evaluate this system by Precison at k method [7].

The remains of this paper are organized as follows. Firstly, Section 2 present the related work on legal texts. Next, we describe the conceptual search and an information retrieval (IR) system implemented on a database of Vietnamese Enterprise legal documents in Section 3. Then, Section 4 presents some experiments on each phase of our system. Finally, conclusion and future work are presented in Section 5.

2 Related works

According to our study of Vietnamese Law, Vietnamese legal documents have been composed carefully by experts, include a large amount of clear concepts and have some specific structures [4]. The language presented in these legal documents is Vietnamese written language which is used officially by the Vietnamese Government to ensure the seriousness, accuracy and consistency of these documents. As experts recommend, a law article regularly be analyzed into logical part annotations including the assumption part, the provision part, and the sanction part [4]. Our view on the recognition of logical structure of Vietnamese Law is closed to the latest research of authors in [1][2]. There might be some differences in the structure of legal documents of each nation but the background, motivation and target is the same.

In [1], the authors divided a law sentence into two parts: a requisite part and an effectuation part. These two parts again are composed of three parts: a topic part, an antecedent part, and a consequent part and there are 4 cases have been made from this composing. The authors presented how to model the task as a sequence learning problem. The model is experimented on an annotated corpus of Japanese national pension law sentences. Then, with the logical part

annotations along with using natural language processing (NLP) techniques, the field of AI and Law – Legal Engineering can address a range of problems aiming to achieve a trustworthy electronic society [3].

While legal documents are always concerned with constraining and controlling social activities using documented norms, they are still lack of ontologies, in comparison with other domains such as medicine, engineering, or psychology [5]. In [5], the authors claimed that legal ontologies are about identifying concepts rather than recognizing any types of knowledge or reasoning roles. Any foundation ontology including notions about agent, actions, time, space ... should be constructed to be able to model and understand legal domain. Although there were some available foundation ontologies such as the CYC upper ontology and the IEEE-Standard Upper Ontology (SUO) but they are not sufficiently covering this domain []. The core ontologies about law can be mentioned such as FOLaw and followed by LRI-core [5][6]. In particular, the LRI core presents three layers of abstraction: the foundation ontology containing general concepts, the core-ontology containing typical concepts for law, and the specific domain-ontology (enterprise, criminal law as case studieds)[5]. Next to LRI core, a legal core ontology called Legal Knowledge Interchange Format (LKIF) is developed in Estrella project, contains basic concepts for law with a considerable number of terms [6].

Until now, there are several approaches to ontology design and development mentioned in [8][6]: top-down approach, bottom-up approach and the hybrid approach called "middle-out". With the logical part annotations and ontology annotations along with using natural language processing techniques, the field of AI and Law can address a range of problems such as translating a legal rule into a logical form, question answering in legal domains, automated knowledge extraction from legal texts [3].

In Vietnam, it is still lack of legal domain ontology since the legal processing has not been started. The search engines on Vietnamese legal documents just simply treat legal documents as normal text. Based on our knowledge, it is the first research on semantic aspects of Vietnamese legal text.

3 A Conceptual Search for Vietnamese Legal Text

Our conceptual search for Vietnamese legal text is proposed based on results of two main tasks: structuring legal texts and building legal domain ontology. This search is deployed as an IR system. We firstly present the overall architecture of IR system in Section 3.1. Then, two main tasks would be presented in Section 3.2 and Section 3.3.

3.1 IR System Architecture

The architecture of proposed IR system has two isolated phases which are shown in Figure 1.

In the first phase, there are two main components as follows:

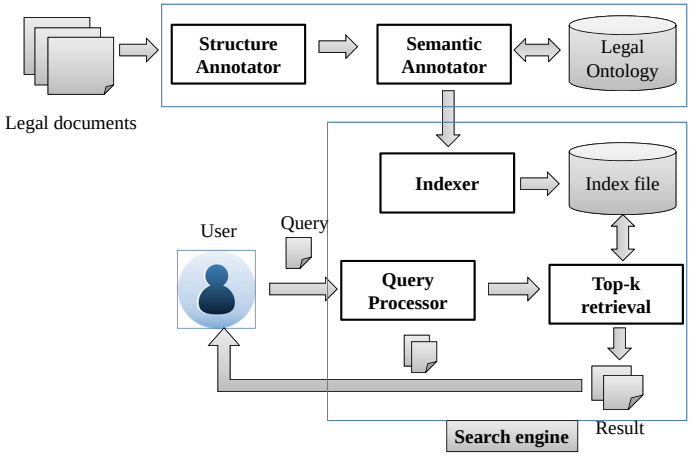

Fig. 1. IR System Architecture

(1) Structure annotator: is built in order to analyze the logical structures of Vietnamese Legal documents stored in law database of Vietnam Ministry of Justice.

(2) Semantic annotator: is consider as one component which takes responsibility for classifying the candidate terms and populating them into high level abstract concepts on our ontology schema, a subtree of LKIF core ontology schema [6][5]. An Enterprise Law ontology is gradually fulfilled and their instances would be matched backwards to documents to annotating them.

In the second phase, a search engine using Apache Lucene libraries [1] will be implemented based on the output of the previous phase. In particular, it would be a collection of annotated documents. An Indexer is used to index this collection and create index files on several fields: full content field, logical structure fields, and semantic concept fields.

The rest of the system is how the Query processor works on the input queries. Each query will be analyzed into a set of keywords (tokens). Next, proximity search method is used to increase the hits of relevant documents. Note that Top-k retrieval method with ranking score computed at Lucene matching model will be used also. A graphics interface is implemented to show out the results to user.

We should say that the core of this system is two main tasks in the first step that are describe below: (1) recognizing of logical structures of Vietnamese legal text; (2) building Vietnamese legal ontology based on LKIF core ontology.

[1] http://lucene.apache.org/

3.2 Recognition of Logical Structure of Vietnamese Legal Text

Our research on Vietnamese legal documents shows that only a few fundamental structures are used. Because legal texts are a kind of highly semantic and consistent document, we decide to apply a rule-based approach to recognize logical parts of text to annotate them based on regular expressions and linguistic features of Vietnamese legal documents.

The structure of legal documents in Vietnamese legal system can be described as follows: each document contains several chapters and each chapter includes at least one article. In some cases, a document contains only articles without any chapters. While analyzing legal articles which are the core of Vietnamese legal documents, we realize that an article is regularly divided into three types of logical part in different orders: the assumption part, the provision part, and the sanction part [4].

In the proposed approach, we study the sample patterns from Vietnam legal documents through analyzing their linguistic features. In particular, we use GATE framework [9] along with some NLP plugins such as ANNIE, VietNLP [2], Hash Gazetteer, and Java annotation patterns engine (JAPE). The input should be a Vietnamese legal text and it would be annotated with several annotation sets: chapter ($< chapter >$), article ($< article >$), assumption ($< a >$), provision ($< p >$), and sanction ($< s >$).

This annotating process is conducted based on the trigger sets which are transformed to hash gazetteers and the patterns written in JAPE transducer. One sample of patterns is presented in Table 1.

Table 1. A sample of pattern in regular expression

Pattern of Recognition
S1= ([<a> **ADV** VietToken* *stopword* | " ,"])? [<a> **AGENT** VietToken*] **PROVISIONTRIGGER** VietToken* ([<a> **ADV** VietToken*])?

The trigger sets includes five kinds of tag as follows:

(1) *adv*: adverbs representing the situations of an assumption part.
(2) *agent*: nouns or noun phrases representing subjects and objects.
(3) $provisionPrefix$: verbs usually appearing in the beginning of a provision part.
(4) $provisionVerb$: main verbs usually appearing in a provision part.

[2] http://vlsp.vietlp.org:8080/demo/?page=resources

(5) *sanctionVerb*: main verbs usually appearing in a sanction part.

Note that mentioned-above *provisionTrigger* is composed of *provisionPrefix* and *provisionVerb*.

The output of these steps is an annotated text with logical structure annotation sets. Some annotated samples of article in Vietnamese and English are described in Table 2.

Table 2. Several samples of Vietnamese legal article (and in English)

Vietnamese	English
<**a**>Đối với Dự án trong Danh mục dự án đã công bố có từ 2 Nhà đầu tư trở lên cùng đăng ký thực hiện</**a**>, <**a**>Cơ quan nhà nước có thẩm quyền</**a**> <**p**> phải tổ chức đấu thầu rộng rãi trong nước hoặc quốc tế để lựa chọn Nhà đầu tư</**p**>.	<**a**> As concerned with the project in the publicized project directory which has been signed up to the development by at least 2 investors</**a**>, <**a**> authorized government offices </**a**> <**p**> have to give the bidding out the domestic or overseas public</**p**>.
<**a**>Tổ chức kinh tế hoạt động dưới danh nghĩa công ty nhà nước</**a**> mà <**a**>không có quyết định thành lập</**a**> thì <**s**>bị đình chỉ hoạt động và bị tịch thu tài sản nộp vào ngân sách nhà nước</**s**>.	<**a**>Business entity that operates under a state organization name </**a**> <**a**>without the establishment licenses </**a**> will <**s**> be suspended and all of its assets will be confiscated to contribute to the State Budget</**s**>.

3.3 Building Vietnamese Legal Ontology on LKIF Core Ontology Schema

The second task in the mentioned-above first phase is semantic annotating on legal concepts with corpus of Vietnamese Enterprise Law articles. This task is done by "middle-out" approach with four following steps: (1) ontology schema partial reuse; (2) chunk extraction; (3) candidate term selection; (4) term classification and population.

Firstly, we reuse a subtree of LKIF core-ontology schema[6] and do partial mapping to concepts in Vietnamese legal text, as described in Figure 2. The concepts at leaf nodes would be populated with terms. Then, the candidate terms which are extracted from the output of automatic linguistic analysis (chunking) will be selected to populated into classes of the ontology schema. This step is repeated to create a fulfilled ontology from Vietnamese Legal Text.

In the initial approach, noun chunks are chosen as candidate terms and we just use 5 concepts as main classes to classify terms. The rule to extract noun phrases is presented (in POS display) in Table 3.

Methods in [10] are used to select candidate terms by calculating phrases' N/NC value. The candidate ones will be classified into five main classes by a

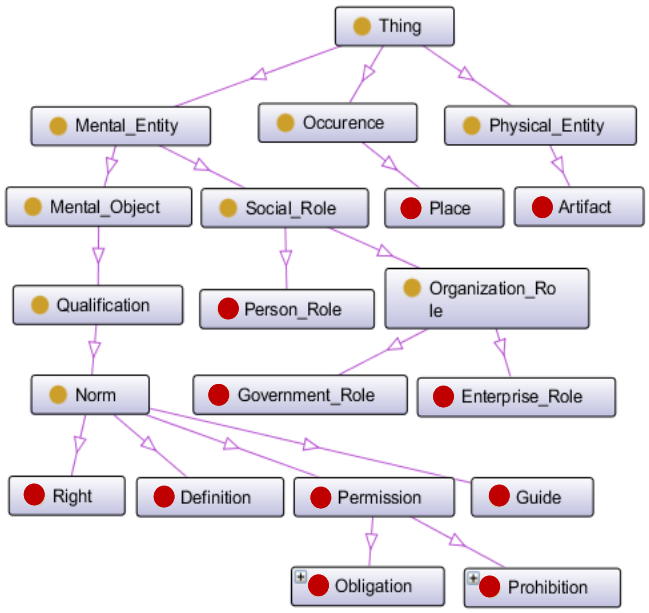

Fig. 2. Partial mapping from LKIF core ontology to Vietnamese Enterprise Law

Table 3. Pattern of Recognizing Noun Chunk

	Pattern of Recognition (in POS display)
Noun Chunk	(DETERMINER)*(NUMBER)*(ADJECTIVE)* <**NOUN**> (ADJECTIVE)* (ADVERB)*

SVM classifier [3]: "Ca nhan" (*Personrole*), "Doanh nghiep" (*Enterpriserole*), "Nha nuoc" (*Governmentrole*), "Tai san" (*Artifact*), and "Noi chon" (*Place*). Finally, we will put legal terms into the ontology schema to form it gradually fulfilled.

4 Corpus and Evaluation Methods

This section presents our corpus for each task belongs to proposed framework and associated evaluation methods.

[3] http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/

4.1 Vietnamese Legal Document Database and Training Corpus

We have implemented a semi-automated tool to collect and purge legal documents from Vietnam Ministry of Justice. A database of 83 Vietnamese Enterprise legal documents including 9379 sentences within 1997 articles is stored sustainably. Each document is stored along with its metadata such as: official number, issued date, validated date, legislation type, effective area, status, expired date, expired part, and altered part.

After that, according to our study of Vietnamese legal text, a golden corpus of annotated Vietnamese Enterprise Law articles is built manually and stored in XML format. Some statistics on this corpus are shown in Table 4. We have some remarks here. About 59.3% of logical part is assumption, 35.1% is provision and only 5.6% is sanction.

Table 4. Statistics on Vietnamese Enterprise Law rule corpus

Type	Number	Part type	Number
Sentence	529		
Rule	132	assumption	393
		provision	233
		sanction	37

4.2 The Task of Recognition of Logical Structure

We directly applied rule-based approach with our own recognition patterns to experiment the corpus. The results were evaluated using precision, recall and $F_{\beta=1}$ scores as follows [11]:

$$precision = \frac{\#correct\ parts}{\#predicted\ parts}$$

$$recall = \frac{\#correct\ parts}{\#actual\ parts}$$

$$F_{\beta=1} = 2 * \frac{precision\ * recall}{precision + recall}$$

A part is recognized correctly if and only if it has correct start word, correct end word and correct annotated type [1].
The result is shown in Table 5.

Table 5. Experiment result of rule-based approach

Annotation	Precision (%)	Recall (%)	$F_{\beta=1}$ (%)
assumption	67.13	61.83	64.37
provision	71.2	58.37	64.15
sanction	86.2	67.57	75.76

4.3 The Task of Ontology Instances Classification

On this task, there are 2733 noun chunks extracted from database of Vietnamese Enterprise Law and there are 299 candidate terms which are selected after that. In this experiment, we implemented SVM method with LIBSVM with a set of 5 linguistic features corresponding to each candidate term: character sequence, previous word, following word, previous part-of-speech, and following part-of-speech.

The training sample set and test sample set are built based on the above candidate terms with mentioned linguistic features. So, the size of training set reaches to 6500 samples. We also make a test set with 500 samples.

After training, the classifier achieved 86.4% in precision score with 432 correct samples over 500 test samples.

4.4 Evaluation on Information Retrieval System

We evaluated our IR system with the mentioned-above database. The range of criteria used for evaluation includes: on the whole document and on logical parts of the legal article.

In the first experiment, we used Precision at k method to identify the relative precision with proximity search function in Lucene framework to evaluate the system. The result is described in Table 6.

We have some remarks here about the correlation between Top k and word distance (in proximity search). The best precision is at case of $k = 30$ and $distance = 1$. In almost case, the precision goes down when distance increases. So, it is not effective with proximity search in Vietnamese legal texts. As table

Table 6. Experiment result in case searching on the whole document

	Top k		
Word distance	K=10 (%)	K=20 (%)	K=30 (%)
1	56.00	**62.14**	**71.07**
2	54.00	59.93	70.00
3	60.00	57.75	59.69
4	**61.00**	58.10	59.05
5	**61.00**	57.18	59.64

shows, when k increases to 30, the higher word distance lead to much worse precision.

In the second experiment, we performed searching on the whole document and on logical parts of the document with associated keywords. For example, a keywords set related to assumption will be used to evaluate system on two ranges of criteria: assumption and full-text. The result in Table 7 describes the percentage of relevant documents over all retrieval documents.

Table 7. Experiment result in various range of criteria

		Keyword set		
		Assumption	*Provision*	*Sanction*
Range of criteria	*Assumption*	80%		
	Provision		63.33%	
	Sanction			64.67%
	Full-document	85%	62.22%	52%

We should say that the result while searching on assumption part got lower precision than the result while searching on full-text. Note that we use a keywords set related to assumption here. The reason is the number of assumption is large but recognition of assumption part just got low precision.

5 Conclusion and Future Work

We presented a conceptual search for Vietnamese Legal text. This search is based on outputs of two main task: recognizing logical structures and building ontology of Vietnamese legal text. We implemented an information retrieval system using this conceptual search to work on documents stored in law database of Vietnam Ministry of Justice. With the input as a database of logical part annotated and semantic annotated legal documents, the system consists two phases that aim to retrieve most relevant documents; firstly, annotating logical parts and semantic concepts; secondly, using Lucene libraries to index on annotation fields and search with proximate range. Experiments on each individual phase showed that the task of recognizing can achieve positive result in $F_{\beta=1}$ score on Vietnamese legal article corpus and the classification model for Vietnamese Enterprise Law ontology got 86.4% in precision. The system is also evaluated by Precision at k method on other ranges of criteria. It is the baseline for future works on legal domain in Vietnam.

However, there are still some limitations in our research such as the small corpus and old methods utilized. Vietnamese legal ontology schema has been built but it is still lack of concepts. To get better result in further research, the corpus and ontology should be extended and state of the art methods should also be implemented. We also plan to extract verb phrases as candidate terms and expand the number of concepts in the ontology schema.

Acknowledgment

This work is supported by Vietnam National University, Ho Chi Minh City. We would like to thank Akira Shimazu and Nguyen Le Minh who has recommended this research topic to our research group at Information System department.

References

1. N. X. Bach, M. L. Nguyen, and A. Shimazu, "Rre task: The task of recognition of requisite part and effectuation part in law sentences," *Int. J. Comput. Proc. Oriental Lang.*, vol. 23, no. 2, pp. 109–130, 2011.
2. M. L. Nguyen, N. X. Bach, and A. Shimazu, "Supervised and semi-supervised sequence learning for recognition of requisite part and effectuation part in law sentences," in *FSMNLP*, 2011, pp. 21–29.
3. M. Nakamura, S. Nobuoka, and A. Shimazu, "Towards translation of legal sentences into logical forms," in *JSAI*, 2007, pp. 349–362.
4. T. N. Duong, N. C. Viet, P. H. Nghi, V. K. Vinh, L. M. Thong, N. V. Dong, and N. V. Huong, *General Theory of State and Law*, 2004.
5. J. Breuker, "The construction and use of ontologies of criminal law in the ecourt european project," in *Proceedings of Means of electronic communication in court administration*, 2003, pp. 15–40.
6. R. Hoekstra, J. Breuker, M. D. Bello, and A. Boer, "Lkif core: Principled ontology development for the legal domain," in *Law, Ontologies and the Semantic Web*, 2009, pp. 21–52.

7. C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge University Press, 2008.
8. E. Francesconi, S. Montemagni, W. Peters, and D. Tiscornia, "Integrating a bottom-up and top-down methodology for building semantic resources for the multilingual legal domain," in *Semantic Processing of Legal Texts*, 2010, pp. 95–121.
9. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "A framework and graphical development environment for robust nlp tools and applications," in *ACL*, 2002, pp. 168–175.
10. K. T. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms: the c-value/nc-value method," *Int. J. on Digital Libraries*, vol. 3, no. 2, pp. 115–130, 2000.
11. C. Goutte and É. Gaussier, "A probabilistic interpretation of precision, recall and *f*-score, with implication for evaluation," in *ECIR*, 2005, pp. 345–359.