Ontology-Based Query Translation for Legislative Information Retrieval

**Kuldar Taveter, Aarno Lehtola, Kristiina Jaaranen,
Juha Sorva, Catherine Bounsaythip**

VTT Information Technology, P. O. Box 1201, FIN-02044 VTT, Finland
Phone: +358 9 456 6044. Fax: +358 9 456 6027.
Email: kuldar.taveter@vtt.fi

http://www.vtt.fi/tte

Abstract

The paper presents a conceptual approach that helps non-expert users to find legislative documents from textual databases. The concepts and inter-concept relationships of each legal domain involved are modeled by an ontology. Terms describing a concept at different levels of accuracy are attached to the concept, and a weight expressing the level of accuracy is associated with each of them. Initial queries presented by the user are matched against these terms. Thereafter the user is shown a graphical representation of the relevant subpart of the ontology that he can use for refining the query. The conceptual approach is preferred over traditional thesaurus because legal terms depend on the differences in law systems that can be expressed by ontologies.

1. Introduction

It is a basic right of citizens to have access to the laws and other legal texts that affect their life. The multi-national legislation of the European Union has made it even more important for ordinary citizens, as well as for professional law experts, to be able to read directives and other legal texts in their native languages.

Recently, it has become easier to get access to different legislative textual databases, but having access does not by itself guarantee that one is capable of finding the exact information needed for the situation at hand because in the countries of EU, directives and legal texts are stored in databases with different structures and in different languages. In addition to the EU legislation, every nation also has its own national laws stored in databases, sometimes even in multiple languages. Finnish laws, for example, are written in Finnish and Swedish as both are official languages in Finland.

Generally, two linguistic phenomena limit the ability of a user to choose effective search terms:

- polysemy (word having multiple meanings): reduces *precision* (the words used to match the relevant document may have other meanings that will be present in irrelevant documents found);
- synonymy (multiple words describing the same concept): reduces *recall* (word-based search needs to match the exact word form).

Legislation is a domain that has many special features deviating from normal everyday texts we read. The use of certain terms in legislative texts differs from general usage of these words, i.e. a word with a common meaning can also have a totally different meaning in the legal terminology. For example, in a law text the word ”consumer” may stand for ”the principal contractor”, among other possible meanings. When one wishes to find legal texts from different databases, it is thus not enough to use a general terminology to define query terms. Moreover, due to differences in law systems, legal terms are not always equal or even compatible in different languages, or even within one language when used in various contexts. For example, a prime minister is called ”premiärminister” or ”statsminister” depending on the country where these Swedish words are used.

Because of the issues mentioned above, some kind of help is necessary in defining query terms and finding the exact information needed for the situation at hand. The user needs not only a list of terms to search with, but also some kind of information about the way these words appear in the legislative texts and the meaning and function they have in them.

2. Representation of legal domains by ontologies

We make use of *ontology models* to give the user more information about legal terms. Ontology is a conceptual model of a problem domain. According to Guarino et al [1], **ontology** can be understood as an intensional semantic structure which encodes the implicit rules constraining the structure of a piece of reality. Ontologies are thus aimed at answering the question “What kinds of objects exist in one or another domain of the real world and how are they interrelated?”. Ontologies can be made explicit by forming a logical theory which gives an explicit and partial account of the above-mentioned intensional semantic structure. Such logical theory contains concepts, their definitions, and relationships between them like e.g. subsumption (inheritance) and aggregation.

We can also say that the use of ontologies in the context of our work is such an extension of the approach of controlled languages described e.g. in [2] where the means the language is controlled by is an ontology.

For specifying ontologies, we utilize the software tool CO(nceptual)N(etwork)E(ditor) [8] which has been developed at VTT Information Technology. Building an ontology for a legal domain with CONE involves:

- creating dictionaries of common concepts from the analysis of existing repositories (law texts) of the law system at hand;
- associating each concept with the terms in the language covered by the ontology;
- assigning to each term the weight expressing the level of accuracy of representing the concept;
- finding and expressing relations between the concepts.

The ontology models in CONE consist of *concepts*, *relations* between them, and *bridges* between the concepts in different models. An ontology model is designed to be used visually, i.e. the user can see a graphical representation of the model. To support the visual effect, concepts of different types are grouped semantically according to their functions.

Following the lines set in the paper [9], the basic concept types in our ontology models of law are *Actors* (humans or institutions), *LegalActs* (actions, activities) that they perform, *LegalInstruments* (documents) that they sign and thereby authorize, *Liabilities* (legal obligations) that they are liable for, and *MainConcepts* and *ConceptParts* that can respectively serve as patients and objects of LegalActs and LegalInstruments, and as consequences of LegalActs. In addition, there are concepts called *Headlines* that are important for visual grouping of the concepts. Headlines themselves are not actual concepts, they just give surface information in the graphical representation of a model. Please note that the type of a concept is often not absolute: the same concept can be seen as e.g. LegalAct or MainConcept, depending on the context and viewpoint. For instance, the concept ”repay” belongs to the type MainConcept in Figure 1, but it could also belong to the type LegalAct.

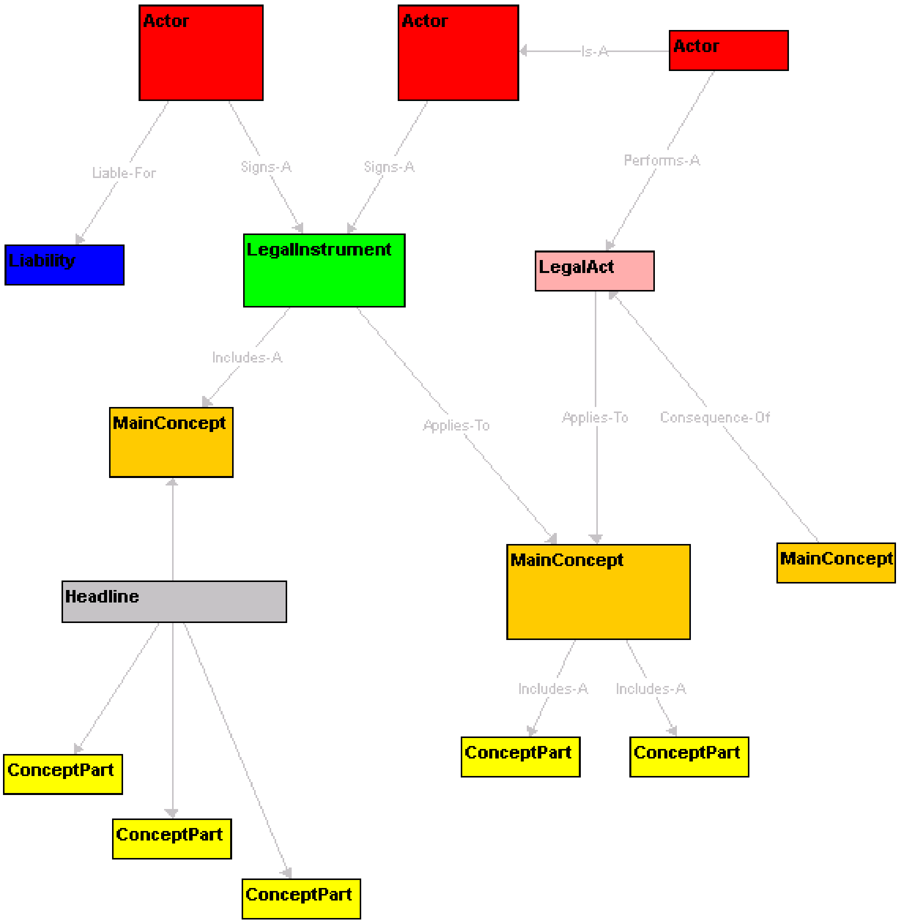


Figure 1. The organization of concepts in the ontology on a travel package law

In Figure 1 the types of concepts and relations in the ontology describing a package travel law is depicted. As it can be seen, concepts of different types are presented with different colors. The figure reveals that a legal ontology is not a traditional tree-formatted hierarchy. A tree representation would require unambigous subsumption relations which are not sufficient enough to describe the complexity of a law in a way that is also understandable to an ordinary citizen.

The inter-concept *relations* shown in Figure 1 reflect semantical relationships between different law concepts. The types of relations used in our legislative ontologies and their meanings are given in Table 1. In the graphical representations of the ontologies, general relations of the type ”Connects” are depicted as arrows without any type labels.

Concepts of different types are divided into clusters according to their functions. For example, the following clusters are present in Figure 4, which is a snapshot of the user interface of our system: ”Acts”, ”Consequences”, ”Consumer”, ”Organiser”, ”Package”, ”Headlines”, and ”Information”.

Each concept of the ontology is represented by natural language expressions called *terms* which can be single words or longer surface expressions. In a concept box of the graphical representation, there is a list of terms from very accurate ones to more general forms, for example the expressions ”package travel contract”, ”package contract”, and ”contract”, assuming that all the expressions do appear in the actual law texts. In the following a concept is referred to by its first term, because it is used most frequently to represent the concept and is therefore at the top of the list of terms of the concept.

According to our approach, each term pertaining to a concept is assigned the value between 0 and 1 expressing the level of accuracy the term represents the concept with. This enables the use of fuzzy logic described e.g. in [3] for matching query terms against concepts. For example, the term ”package travel contract” representing the concept of the same name is assigned the value 1.0, and the terms ”package contract” and ”contract” pertaining to the same concept are assigned the values 0.7 and 0.3, respectively. The use of accuracy values is explained in section 3.

RELATION	MEANING
Connects	General (unspecified) relation
Is-A	Subsumption
Includes-A	Aggregation
Performs-A	Actor → Legal Act
Signs-A	Actor → Legal Instrument
Applies-To	Legal Act → Patient, Legal Instrument → Object
Consequence-Of	Product of the Legal Act → Legal Act
Agent-Of	Representative of the Actor → Actor
Role-Of	Role of the Actor → Actor

Table 1. The types of relations and their meanings

The links between different ontological law models are organized with *bridges* that lead from one ontology model to another and thus guide the translation between the models and languages. There can be a bridge between any two concepts in two different ontology models. Usually there exists a bridge between concepts that are translational equivalents in two languages and/or share the same function in different laws. According to [4], if a concept does not have any kind of counterpart in another ontology model, it can be linked to a superclass concept in the target model. For example, since the concept ”varallisuusvahinko” (damage done to one’s property) in Figure 2 has no equivalent concept in the target model, it can be connected to the concept ”damage” at the superclass level. Even if the bridge does not describe an equivalent translation, the user can understand the connections between the concepts by visual comparison of the graphical representations of the original ontology model and the target model.

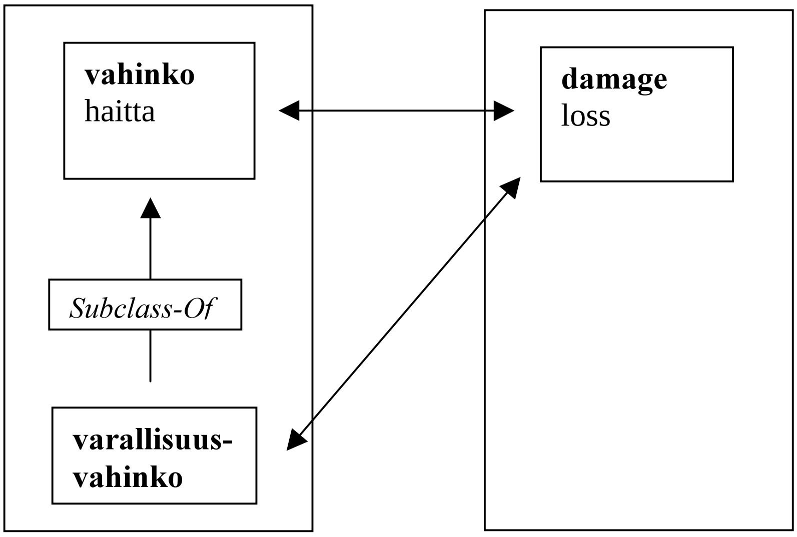
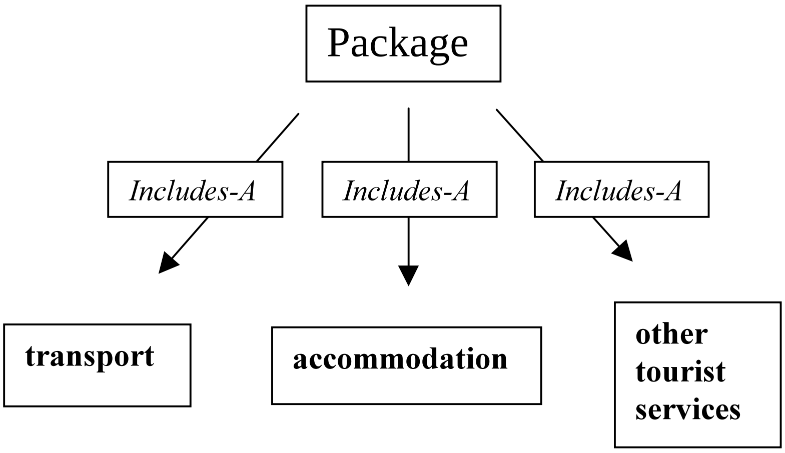


Figure 2. Formation of bridges between concepts of different ontologies

Let us take a simple example of conceptual modeling of legislation by an ontology. The package travel law of EU defines the concept ”package” as follows:

"package" means the pre-arranged combination of not fewer than two of the following when sold or offered for sale at an inclusive price and when the service covers a period of more than twenty-four hours or includes overnight accommodation:
(a) transport;
(b) accommodation;
(c) other tourist services not ancillary to transport or accommodation and accounting for a significant proportion of the package.

A piece of an ontology model based on the previous law text could be:



The concept ”package” consists of different parts which are represented in the model above. The relation between the concepts is ”Includes-A” (aggregation). Other terms referring to the concept ”package” are then added, and the values describing the levels of accuracy of representing the concept are assigned to them. After that the concept could look like this:

Package
package: 1.0
package travel: 0.9
package holiday: 0.7
package tour: 0.5
inclusive tour: 0.3

The concept ”package” is then linked to a comparable concept in a model describing another law system in another language (for example to the Swedish model):

Package
package: 1.0
package travel: 0.9
package holiday: 0.7
package tour: 0.5
inclusive tour: 0.3

↔

Paketresa
paketresa: 1.0
resepaket: 0.9
semesterpaket: 0.7
paketarrangemang: 0.5

Please note that the number of terms referring to the same concept can vary in different languages. Please note also that the expressive power of our present model does not enable to represent the *rule* present in the definition of the legal term ”package”, according to which ”package” means the combination of *at least two* of the following: transport, accommodation, and other tourist services. Expressing of this rule is, however, important for expressing legislative norms *per se*, but not to translating between the concepts of different law systems expressed as ontologies.

3. Ontology-based query matching and translation

Our system is embedded into an interface through which the user can make queries in one language to search for legislative texts from different EU and national databases. By using that interface, the user selects the source and target languages and systems of law (e.g. EU or national). Based on them, the source and target ontologies are determined. The first step in the ontology-based query translation is matching query terms against terms representing concepts in the source ontology in order to determine the concept most probably relevant to the query. The values expressing the levels of accuracy of terms in the source ontology are used for that. For example, let us assume that the source ontology is an ontology about the EU legislation on package travels expressed in English. If the query term used is ”supplier”, it can be matched with two concepts of the ontology: ”retailer” and ”organizer”, because they both include ”supplier” as one of their terms[1]. In order to decide between the two alternatives, accuracy values and fuzzy logic are used as depicted in Figure 3. Since the accuracy values representing the degrees with which the term ”supplier” represents the concepts ”retailer” and ”organizer” are respectively 0.6 and 0.2, it can be concluded that the degree of membership of ”supplier” in the concept ”retailer” (60%) is higher than the degree of membership of ”supplier” in the concept ”organizer” (20%), and the concept ”retailer” can be assumed to be more relevant than the other one. The number of alternatives to decide between can naturally be even higher than two.

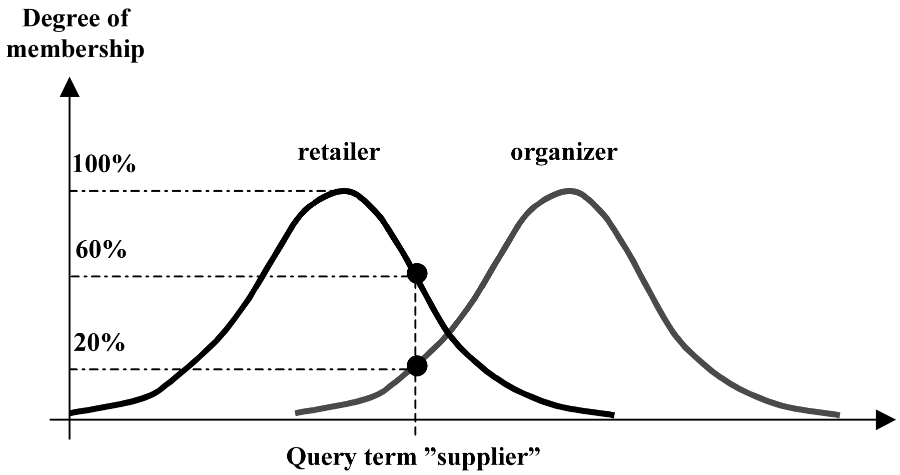


Figure 3. Comparing the term’s degrees of membership in two concepts

At the next step of ontology-based query translation, the source ontology is represented graphically with the found most relevant concept highlighted. The described situation is depicted in Figure 3 as a snapshot of the user interface of our system. In the upper rightmost window of the figure, the English terms of the EU legislation on package travels representing the concepts initially found from the query or selected by the user are presented. The lower window in the right shows the corresponding terms in the legal terminology of the Finnish national law on package travels. The middle window depicts all the terms of the Finnish national law on package travels. From there the user proficient in Finnish can directly select the terms he is interested in, thus skipping the phase of browsing the ontology. The translation between the laws of EU and Finland is performed by using the ”bridge” relationships (cf. Section 2) between the concepts of ontologies about EU’s and Finnish legislation on package travels. The user can view the latter ontology by pushing the button ”Display target model” depicted in the figure.

Since the source ontology also presents other concepts related to the original target of the search, the user can *expand* the search by choosing other related concepts in the ontology model. For example, in Figure 4 the user has also selected the concept ”local agencies” which is linked to the concept ”retailer” by the relation of the ”Agent-Of” type, as a result of which the translated query will be expanded with the terms representing the concept ”local agencies” in the target language, i.e. in Finnish. The accuracy values can also be used for weighing translated queries by assigning each

[1] A term representing the concept ”organizer” is actually ”service supplier”, but according to the dissertation [7], basic words of phrases like e.g. ”supplier” of ”service supplier” can be used as terms with reduced weights (accuracy values)

component term of a translated query a weight proportional to its accuracy value in the target ontology. Moreover, each relation type can also be assigned a similar accuracy value representing the strength or reliabilty of the relation. For example, the strengths of the ”Is-A” and ”Applies-To” relations could be 0.8 and 0.5, respectively. These values can be used for additional expansion and weighing of search terms. A relevant methodology is described in the work [7]. Consequently, our approach also provides *metadata* that can be used for weighed query expansion in accessing legislative textual databases. However, at present our system is not used for that.

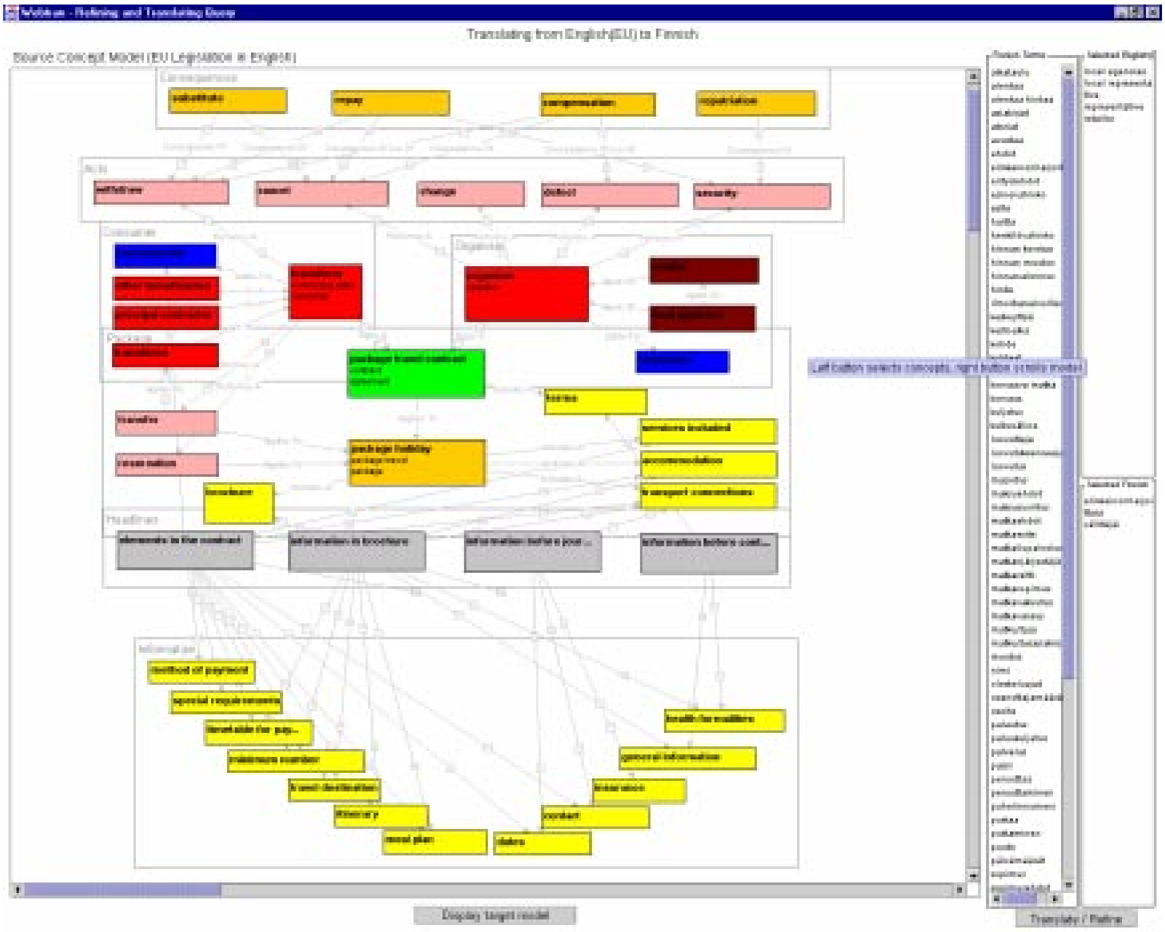


Figure 4. Ontology-based translation of terms from the EU legislation on package travels expressed in English to the Finnish national legislation on package travels

The law systems and languages that can be translated between with the help of our system are currently:

- the legislation of EU in English, Finnish, Swedish, and French;
- the national legislations of United Kingdom, Finland (in Finnish and Swedish), Sweden, and France.

Presently the only subarea of law that is supported is the legislation on package travels, but other subareas will be added in the near future.

Our system has been implemented as a Java applet which has an interface to the search engines connected to legislative databases. The query expressions, as well as the choices of the language and law system by the user, are passed to the applet through HTML applet parameter tags as character strings.

The applet connects to the serverside or local relational database of ontology models, loads the relevant models, and displays them to the user within the WWW browser that is used to view the applet.

4. Related work

Our method of representing ontologies is based on *conceptual graphs* which is a knowledge representation language introduced by John F. Sowa [12]. Conceptual graphs use graphical representation as a method of encoding knowledge. In a conceptual graph, concept nodes are used to represent entities, attributes, states, and events, while relation nodes are used to show how these concepts are related to each other. One of the main differences between our ontology models and conceptual graphs is that in our models attributes of a concept are represented as properties of the concept, while in conceptual graphs attributes are represented as concept nodes.

The Hyperlaw system described in [5] is based on an information retrieval system model called EXPLICIT. This model uses a two-level structure to make the different parts constituting the body of data explicit at the document level, called the hyperdocument, and at the auxiliary data level, that is, at the level of the semantic structure according to which the indexing terms are organized, called the hyperconcept.

The two levels are linked by the relations between the concepts and the documents the concepts describe. At the same time the single elements contained on each of the two levels are interconnected: the documents are linked by references or citations, while the links among the auxiliary data are made up of the semantic structure in which the terms are placed.

The EXPLICIT model makes it possible to display the conceptual structure of the indexing terms, so that the proper semantic context in which each term is placed becomes known according to the meaning it is given in the indexing phase. This semantic structure can, at all times, be actively searched by the user, who, therefore is able effectively and immediately to enhance his query.

The model makes use of a conceptual framework tailored for a specific application domain and makes that scheme available for active utilization by the user during his interaction with the system, thus providing a reference structure for the process of query formulation. Another feature is concurrent use of different conceptual schemes for the same application domain.

The FOOD system described in [6] proposes a Fuzzy Object Oriented Data Model representing both vague and uncertain information by means of linguistic qualifiers. The interpretation proposed for uncertainty qualifiers is based on the assumption that they contain an implicit twofold information: they both specify that the declared (vague) attribute value can be violated by the actual value to a given degree, and impose an imperative safeguard constraint on this possible violation.

5. Future work

The present version of our system concentrates only on conceptualizing and visualizing package travel law, a very specific subdomain of legislation. In the future, as more models of other subdomains will be created, we will be faced with the problem of bringing together these parts to form an easy-to-use system for searching the ontology models pertaining to many fields of legislation.

Obviously, as the number of legislative fields increases, the ontology models will fast become much too numerous to be incorporated into a single, unified model. Moreover, many of the legal terms involved might carry different connotations (and thus have different conceptual connections) in different subdomains, resulting in ambiguity, and possibly considerable confusion for the end user. Therefore hierarchical organization of the legislative fields is necessary in order to isolate the various subdomains and provide the user with a sensible method for accessing the correct one.

In the hierarchical domain system, the ontology models would be divided into several layers, which branch into the layer beneath them. The top level would include the very basics of legislation – the concepts that describe the law-making process in general and define the major subdomains of the legal system in question (e.g. customer protection law, criminal law, etc.). Each of these subdomains would in turn contain concepts corresponding to sub-subdomains, which the user could select and view. For example, the user could select the concept "customer protection law" at the top level and then move on down to "travel law" and finally to "package travel law", if he wanted to find the very basic concepts involved in these specific laws.

In the later versions of our system, the end user would be able to navigate graphically through the described layers. The concepts in the top-level model would correspond to the subdomains of the top ontology model, and by selecting the desired concept the user would be able to zoom in and view the second layer of the hierarchy, with more detailed concepts. At the bottom level, the user could inspect the basic concepts and terms of one particular subdomain (e.g. legislation on package travels).

There would also be implemented an automatic search system for the ontology models of the various subdomains, which would, if wished, automatically direct the user to the relevant subdomain, based on the query submitted by the user. In case of ambiguities (i.e. the same term is used in different contexts in different subdomains), fuzzy matching would be used. As another alternative, the user would be taken to the lowest possible level containing all the subdomains in question. For example, if the user made a query with a term used both in the legislation on package travels and product marketing, he would be shown the "customer protection" ontology model with the concepts "package travel law" and "product marketing law" highlighted.

As another important issue to be treated in our future work, a well-defined methodology for capturing the concepts, relations between them, and the terms representing the concepts should be worked out. The methdology should also provide guidelines for assigning accuracy values to terms and possibly also to relations, as this is presently done just by intuition.

Our approach can also be applied in domains other than legislative ones. The use of a similar methodology in the domain of foreign trade is described in the paper [4].

Our work should also be placed into the framework of task/context analysis as described e.g. in the papers [10] and [11].

It is also worthwile to consider how could new terms be included at run time, so that the ontology model and the query at hand would adapt themselves to these terms.

6. Conclusions

The *main contribution* of our approach is that instead of using a general language in a search for legislative documents, the user can take advantage of the legal terms provided by the ontology model. The user starts out with general-language or specific query terms that lead to certain concept(s) in the source ontology representing the legal system and language selected by the user. As another novel feature of our approach, *fuzzy logic* is used for matching query terms against terms representing concepts in the source ontology in order to determine the concept(s) relevant to the query.

In the source ontology, the user can specify the *exact concepts* to be used in the query to the databases. Since the ontology also presents other concepts related to the original goal of the search, the user can *expand the search* by choosing related concepts in the ontology model. The terms representing the selected concepts are then translated according to the links between the source ontology and the ontology model of the target law system and language.

Our approach also provides *metadata* that can be used for *automatical expansion of translated queries and weighing of their constituent query terms* in accessing legislative textual databases.

The ontology models can also be of advantage in the actual legislative process, for example when *harmonizing the terminology* of legal areas between different countries and languages. This could offer considerable benefits for example in applying the EU directives to national laws. Even the information structures of law texts could be harmonized through the use of the ontology models describing the contents of the laws.

The harmonizing process could be based on the *intralingual translation* for example between the terminology used in the British legislation and the terminology of the EU-directives in English. During the design of the ontologies for our system it was noticed that the terminology in Finnish had also considerable discrepancies between the national and EU legislation. The ontology models can offer the user information about the terminology in one specific language used in different legal sources.

7. Acknowledgements

The described work is mainly funded by the Technical Development Centre of Finland (TEKES). The authors are very grateful to TEKES and to the industrial partner of the project TietoEnator Group from Finland.

8. References

1. Guarino, N., Giaretta, P. Ontologies and Knowledge Bases: Towards a Terminological Clarification. In: N. J. I. Mars (ed.), Towards Very Large Knowledge Bases, IOS Press 1995, pp. 25-32.
2. Lehtola, A., Tenni, J., Bounsaythip, C., Jaaranen, K. WEBTRAN: A Controlled Language Machine Translation System for Building Multilingual Services on Internet. Machine Translation Summit VII ‘99 (MT Summit 99), Sep. 13-17, 1999, Singapore, 9 p.
3. Cox, E. The Fuzzy Systems Handbook. 2nd ed., AP Professional, 1998.
4. Taveter, K. Intelligent Information Retrieval Based on Interconnected Concepts and Classes of Retrieval Domains, Proceedings of the Eigth DELOS Workshop User Interface in Digital Libraries, Stockholm, Sweden, 21-23 October, pp.39-43.
5. Agosti M., Colotti R., Gradenigo G. A Two-level Hypertext Retrieval Model for Legal Data. In: A. Bookstein, Y. Chiaramella, G. Salton and V.V. Raghavan (eds.), Proceedings of the 14th ACM-SIGIR International Conference on Research and Development in Information Retrieval, Chicago, USA, 1991, 316-325.
6. Bordogna, G., Pasi, G. Linguistic Qualifiers of Vagueness and Uncertainty in a Fuzzy Object Oriented Data Model, Proceedings of the Third International ICSC Symposia on Intelligent Industrial Automation (IIA’99) and Soft Computing (SOCO’99), Genova, Italy, June 1–4, 1999, pp. 719-725.
7. Kekäläinen, J. The Effects of Query Complexity, Expansion and Structure on Retrieval Performance in Probabilistic Text Retrieval. Academic Dissertation, University of Tampere, Department of Information Studies, Finland.
8. Kankaanpää, T. Design and Implementation of a Conceptual Network and Ontology Editor. VTT Information Technology, Research Report TTE1-4-99, June 1999, 74 p.
9. Hafner, C. Representation of knowledge in a legal information retrieval system. In: Information Retrieval Research, R. Oddy, S. Robertson, C. van Rijsbergen, P. Williams (Eds.), Buttersworth, London, 1981.
10. Holtzblatt, K., Beyer, H. Contextual Design: Principles and Practice. In: Field Methods for Software and Systems Design, D. Wixon and J. Ramey (Eds.), John Wiley & Sons, Inc., New York, 1998.
11. Richardson, J., Ormerod, T. C., Shepherd, A. The role of task analysis in capturing requirements for interface design. Interacting with Computers 9(1998), pp. 367 – 384.
12. Sowa, J. F. Conceptual Structures, Information Processing in Mind and Machine. Addison-Wesley Publishing Company, 481 p.