Information Extraction based Multiple-Category Document Classification for the Global Legal Information Network

Richard D. Holowczak **Nabil R. Adam**
Rutgers University, Center for Information Management,
Integration and Connectivity (CIMIC)
180 University Ave., Newark, NJ 07102
{holowcza,adam}@cimic.rutgers.edu

Abstract

This paper describes a prototype application of an information extraction (IE) based document classification system in the international law domain. IE is used to determine if a set of concepts for a class are present in a document. The syntactic and semantic constraints that must be satisfied to make this determination are derived automatically from a training corpus. A collection of IE systems are arranged in a classification hierarchy and novel documents are guided down the hierarchy based on the results from the previous level. Experimental results for a research prototype are given on a subset of the Global Legal Information Network domain.

Introduction

The Global Legal Information Network (GLIN) is a database of international laws maintained by the U.S. Law Library of Congress (LLoC) (Adam *et al.* 1996). Member countries (about 35 and growing) from around the world submit statutes and regulations in their native languages. Law experts at LLoC summarize the incoming laws in English and assign index terms to the summary from a controlled legal thesaurus. Laws can later be retrieved by specifying one or more index terms.

The summarization and index term assignment activity is labor intensive and requires law specialists who are familiar with the national language of origin. Several forces are at work to make this task even more daunting. The number of countries submitting laws is expected to increase sharply in the next several years while the types of laws may be expanded to include case law and additional types of statutes. In addition, demand for the accurate and timely retrieval of relevant legislation has driven the need for uniformity in assigning the index terms. Our application of document classification to GLIN aims to provide quality control of the index term assignment process (essentially a classification task) by providing a parallel path through the classification process. Thus our goal is to support, not supplant, the efforts of the law experts.

Document classification is the process of assigning a document to one or more classes (Salton 1989). It requires the definition of a set of target classes and algorithms that take documents as input and assign them to one or more classes. The majority of existing approaches use the frequency of words in documents as evidence towards making a class assignment. Other approaches use hand-coded rule bases, case frames or statistics to attempt to map document characteristics such as word co-occurrences to classes.

In this paper, we present a novel multiple-category classification method based on information extraction techniques. We present a methodology for constructing a hierarchy of concepts that are common to a collection of documents. Such an approach provides a scalable, high precision means to assign text documents to one or more pre-defined classes by using semantic and syntactic sentence analysis. A prototype application to the GLIN law domain is discussed.

Creating a Classification System

We have taken an Information Extraction (IE) approach to the classification problem. Information Extraction is a type of Natural Language Processing (NLP) with the goal of extracting a set of facts or concepts from text documents. Our classification methodology uses the presence or absence of concepts in a document to determine membership in a given class. By concepts, we mean the abstractions or notions one expects to find in all documents in a given class.

The two main tasks in developing an IE based classification system are to model the domain as a conceptual hierarchy and to train the extractor's concept node definitions. These are discussed next.

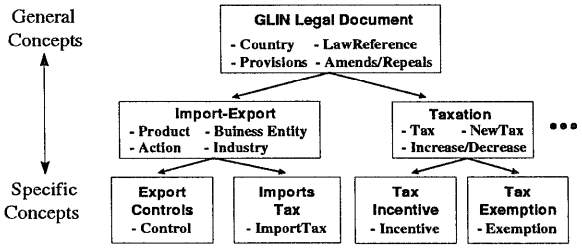

Figure 1: Example hierarchy from the GLIN domain

Defining the Conceptual Hierarchy

A conceptual hierarchy covering concepts within a domain will be used to classify documents. Each node represents a class in the conceptual hierarchy and contains a set of concepts common to all documents within that class. For example, in Figure 1, documents in the `Import-Export` node all have in common the concepts *Business Entity*, *Action*, *Industry* and *Product.*

In general, nodes towards the top or root of the hierarchy contain more general concepts that are common to most or all of the documents while nodes towards the bottom or leaves of the hierarchy contain more specific concepts unique to those classes.

Determining the appropriate concepts for a given node has traditionally been an ad-hoc process achieved through consensus of domain experts. Some tools and approaches to facilitate porting IE systems to new domains are discussed in (McCarthy 1993). The creation of knowledge bases and ontologies are two such areas that could benefit from a structured means of determining "what's important" in a given domain. Some guidelines for what a methodology for constructing ontologies might consist of are given in (Uschold & King 1995).

We employ a methodology (Holowczak 1997) that takes, as a starting point, an existing classification framework that is then filled in with concepts present in the representative documents. We note that there are many domains where such an assumption holds. Consider, for example, the Library of Congress subject headings, the UMLS medical MetaThesaurus or, as in our examples, the GLIN index term thesaurus. In addition, we note that such frameworks are hierarchically arranged although this is not a requirement for the methodology. Another definition for such an arrangement is a taxonomy for the domain where the "is-a" relationship holds between child and parent (e.g., Imports Tax law is-a Import-Export law).

The methodology uses a "bottom-up" approach to identify the concepts and operates on a set of training documents representative of the domain. The training documents are grouped according to their classification which forms the framework for resulting conceptual hierarchy. The methodology is outlined as follows (details of step 2 are presented after the outline for clarity):

1. Assume that the classification hierarchy is represented as a tree with a set of nodes, each having two indexes: l and s. A node is located at a certain level l (where $l = 0, 1, \ldots, n$, with $l = n$ representing the root level and $l = 0$ representing the leaf level. At a given level, nodes are numbered from 0 to m, where 0 is the left-most node.

2. Begin at the left-most leaf node (i.e., $l = 0$ and $s = 0$) and determine the concepts present in documents grouped in this node. Label this set of concepts C_{ls}. This step is only performed for leaf nodes in the hierarchy.

3. Move to a sibling node at the same level l and repeat the previous step until concepts have been determined for all sibling nodes. Label these sets of concepts accordingly.

4. Examine sets C_{00} through C_{0n} where n is the number of sibling nodes. Collect all concepts common across C_1 through C_n (the intersection of C_1 through C_n). Label this set CC.

5. Remove all concepts present in CC from each of the siblings (C_1 through C_n) and assign CC to the parent node.

These steps are then repeated for each of the leaf nodes. Once the leaf nodes have been processed, the algorithm continues in a depth-first fashion by finding common concepts among parent nodes, grandparent nodes, etc. on up the root.

Step 2 requires us to identify all of the concepts in the current leaf node. This task is performed by domain experts who are asked to reach consensus on the common concepts. This approach is valid for our methodology, however, we note that duplication of effort will occur. Identifying concepts for siblings may require the same concepts to be defined multiple times. For example, while collecting C_{i+1}, we will duplicate some effort required to create C_i.

We are presently investigating an automated approach to discovering common concepts among documents within a class is based on work done at University of Utah (Riloff & Shoen 1995). Using the example hierarchy in Figure 1 we provide the following illustration of the conceptual hierarchy modeling methodology.

1. Starting at the left-most sibling "Export Controls", we examine a collection of GLIN summaries that have been assigned that index term (See Figure 2 for an example). We note that each of these summaries has in common a source country, a reference to a decree, statute or other law reference and a number of provisions. Most also contain some indication of other laws that are either amended or repealed by this law. Each of the laws mention some action such as importing or exporting particular products. These actions are carried out by various business entities in a number of industries. Finally, we note that some form of export control is being exerted on the industry, business entity, products or actions in question.

2. This process is repeated for the next sibling node representing the "Imports Tax" class. Most of the same concepts are found in Imports Tax documents: Country, Law Reference, Provisions, Amends/Repeals, Products, Action, Business Entity and Industry. In addition, the concept of a tariff, levy or tax placed on imported products is also present in these summaries.

3. The two sets of concepts from "Export Controls" and "Imports Tax" intersect on Country, Law Reference, Provisions, Amends/Repeals, Products, Action, Business Entity and Industry. Thus these concepts are removed from the two sibling nodes and assigned to the "Import- Export" node.

4. In a similar fashion, the concepts in "Tax Incentive" and "Tax Exemption" are identified and the common concepts are removed and assigned to the "Taxation" node.

5. With the leaf nodes completed, attention is then shifted to the next highest level. after steps 1 through 4, the "Taxation" and "Import-export" nodes each have been assigned a set of concepts each. The concepts in common between the two, in this case: Country, Law Reference, Provisions, and Amends/Repeals. These concepts are removed from the "Taxation" and "Import-export" nodes and assigned to the root node.

Classifier Training

At this point, we have constructed a conceptual hierarchy that mirrors a classification. To train the extractor for a given class, we make use of a training corpus of documents representative of the class. The training involves inducing a set of concept definitions for each extractor. The set of definitions, called a *concept node* is used to determine if a document contains the relevant concepts for a class. For this task, we employ the CRYSTAL concept dictionary induction tool developed by the University of Massachusetts (Soderland *et al.* 1995).

Spain
Source: Boletin oficial del Estado June 22, 1990
SUMMARY
Resolution of June 12, 1990 of the Ministry of Economics and Finance's General Administration of Foreign Commerce provides supplementary provisions concerning the standards of quality for olives for table use intended for export, for 1990, including weights and measures. Includes special provisions for exports to the United States, Puerto Rico, and Canada. These include export restrictions and packaging and labeling requirements. (2 provisions, including tables. Pages 17550-17551.)

Figure 2: Example Export Controls Summary

The training corpus must be prepared as follows. First, a part of speech lexicon is created for the domain. This step consists of augmenting a core set (about 3000) of domain independent word/POS pairs with words found in the training corpus for the class. Some tags from the GLIN domain include nouns like *incentive, imports, law* and *amendment*, verbs such as *amend, repeal* and *abrogate*, and words that appear as both nouns and verbs such as *export* and *decree.* Next, a semantic features list is created with domain specific words and their associated classification in the domain hierarchy. For example, *Venezuela* is associated with `ws_Country`, *Accord* is associated with `ws_LawReference`, *Approval* is associated with `ws_ExportControls`, etc. The general approach used in the first two steps is to order the words in the training corpus by frequency. High-frequency words are then added to the POS lexicon and semantic features list. Last, the training corpus for the class is marked up with SGML style markup tags corresponding to the concepts in the domain hierarchy. An example tagged training text is given in Figure 3.

Inducing Concept Nodes

At this point, the conceptual hierarchy has been modeled, domain specific part of speech and semantic lexicons have been constructed and a set of training documents for each class have been tagged with semantic features.

A concept node definition is a set of semantic and syntactic constraints used to determine the presence of a concept in text documents. Semantic constraints

<CO> *Cape Verde* </CO>
Source: Boletim oficial October 17, 1987
SUMMARY
<LR>*Resolution*</LR> *57/87 of 10/17/87 approves the schedule of* <TX>*taxes*</TX> *for animal inspection and inspection of* <PR>*food products*</PR>*, both of animal and nonanimal origins, which are destined for* <AC>*import*</AC> *or* <AC>*export*</AC>*.*
<RP>*Repeals*</RP> *Table A annex to* <LR>*resolution*</LR> *of* <PR>*livestock*</PR> *Health, approved by Legislative Diploma 1278 of 3/17/56. (3* <LR>*provisions*</LR>*)*

Figure 3: Example tagged training text

CN-type imports-tax ID: 1365 Status: generalized
Constraints:

VERB::	classes:	ws_Root_Class
	mod class:	ws_Root_Class
	head class:	ws_Root_Class
PP::	terms: OF	
	mod terms: OF	
	classes:	ws_GLIN_Class
	mod class:	ws_Root_Class
	head class:	ws_Root_Class
SUBJ::		==> TaxType
	terms: TAXES	
	classes:	ws_ImportsTax
	mod class:	ws_Root_Class
	head class:	ws_Root_Class

Figure 4: An example Concept Node definition for Imports Tax documents

restrict word uses to a given location in the conceptual hierarchy. For example, a possible constraint may be the use of the verb EXPORT in the context of `ws_Activity` only. A syntactic constraint restricts the use of a word in part of speech, active or passive voice (for verbs), location within a particular type of phrase such as a verb phrase or noun phrase, or location within the object or subject of the sentence.

To induce the concept node definitions, the MARMOT pre-processor and heuristic parser is used to parse the training texts. The pre-processor folds the text to upper case, normalizes dates and date references, performs substitution for common phrases (`"DOMINICAN REPUBLIC" -> "DOMINICAN_REPUBLIC"`), identifies punctuation and sentence boundaries, identifies phrases (NP, VP, PP, ADVP) and identifies subjects and objects in sentences.

The parsed texts are then fed to the CRYSTAL concept dictionary induction tool to induce the CN definitions. The overall approach CRYSTAL takes is to begin with an initial set of definitions with word-for-word constraints on all of the words in a phrase where key words have been tagged. This initial set is then iteratively processed to merge and generalize similar definitions. The operation of CRYSTAL is described in depth in (Soderland *et al.* 1995).

The result of this process is a set of concept node definitions that constrain the syntactic and semantic conditions under which concepts for the given class may appear. This set can then be used to determine if a given text document contains these concepts. An example concept node definition for Imports Tax documents is shown in Figure 4. In this example, there are no constraints on the verb, the prepositional phrase must contain the word "OF" and the subject of the sentence must contain the word "TAXES" from the `ws_ImportsTax` class. The sentence: "... approves the schedule of taxes for animal inspection and inspection of food products..." is one example that will satisfy these constraints.

Classification of a Novel Document

Novel documents are parsed by the MARMOT pre-processor/parser and fed as input into the domain hierarchy. The classifier at the root of the domain hierarchy (in this example the `GLIN Legal Document` node) attempts to extract concepts from the document by applying CN definitions using the BADGER sentence analyzer. If no matching concepts are found, the document is classified as outside of the domain. If concepts are found in the document, it is passed from the root node to the child nodes of the hierarchy. For the example text in Figure 3, the concepts `LawReference`, `Country` and `Repeals` are extracted, thus the document is passed on to the child nodes `Taxation` and `Import-Export`.

For the `Import-Export` node, the `Product` and `Action` concepts are extracted. The `Business Entity` concept is not found in this document as it is not expressly stated who will do the inspections or the importing or exporting of products. Following this branch further, the concept of `Export Controls` is not found in the document so that path is terminated. However, at the `Imports Tax` node, we are able to extract a `TaxType` concept. For the `Taxation` node, we are able to extract the concept of `TaxType`. However, we are not able to extract any concepts from the `Tax Incentive` and `Tax Exemption` nodes.

At the end of processing, the IE based classification system has positive results (appropriate concepts were found) for the *GLIN Law Document*, *Taxation*, *Import-Export*, and *Imports Tax* nodes. Based on this output, we can then classify this document by assigning the appropriate classification terms for each of these nodes. For example, the appropriate terms from the GLIN thesaurus would be `Import-Export`, `Taxation` and `Imports Tax`.

Classification Experimental Results

In this section, we present some experimental results for the classification task on a portion of the GLIN domain. These initial tests were done on each individual class. The experimental methodology for each class was as follows:

1. An extractor for the class was trained using 80% of the existing classified GLIN summaries. These summaries had been previously classified by the GLIN law experts.

2. A *test set* of summaries was created comprised of the remaining 20% of the summaries not used for training (relevant summaries) and an equal number of summaries randomly chosen from outside of the class (irrelevant summaries)

3. The test set was then run through the appropriate classifier. The extractor attempted to identify those summaries that should belong to the class and those that should not.

Recall was calculated as the number of relevant summaries correctly classified, divided by the total number of relevant summaries. Precision was calculated as the number of relevant summaries correctly classified, divided by the total number of summaries (relevant and irrelevant) classified.

For example, in the value-added tax class, the test set was made up of 53 relevant summaries (20% of a total of 264 value-added tax summaries) and 53 irrelevant summaries chosen randomly from outside of the value-added tax class. The classifier correctly identified 48 of the relevant summaries (out of a possible 53) giving a recall score of 91%. A total of 49 summaries were identified as relevant with only one being incorrectly included in the relevant group giving a precision score of 98%. Additional scores are given in Table 1.

Class	Recall	Precision
Taxation	87%	90%
Capital Gains Tax	92%	92%
Consumption Tax	95%	95%
Internal Tax	73%	100%
Profits Tax	97%	80%
Stamp Tax	94%	99%
Tax Credits	100%	90%
Travel Tax	100%	100%
Value Added Tax	91%	98%
Import-Export	70%	80%
Export Controls	94%	63%
Export Incentives	97%	67%
Imports Tax	97%	92%

Table 1: Summary of experimental results

Examination of the experimental results reveals several classes, such as Consumption Tax, exhibiting high recall and precision while others such as Import-Export exhibit lower performance scores. High performance can typically be attributed to homogeneity of the phrasing used in the training texts. For example, in the Consumption Tax class, there are only 19 concept node definitions (sets of syntactic and semantic constraints) generated during the training phase. This indicates that the wording used to describe the concept of a Consumption Tax is fairly standard across the training texts. High performance in the experiment indicates that the homogeneity extends to the test documents as well. In the case of lower performing classes such as Import-Export, there are typically many more concept node definitions generated indicating the relative heterogeneity of phrases among the training and testing documents.

We are currently investigating ways to improve the classifier performance. We feel there is a relationship between the accuracy of the model and the performance of the classifier and are thus exploring approaches to develop more accurate domain models. In addition, in the training phase, the CRYSTAL tool has several parameters that can control the degree to which concept node definitions are generalized. Concept node generalization has a direct impact on the recall/precision tradeoff.

Relevant Classification Work

Automated classification approaches differ in two main dimensions. First, the features of the document used to determine the appropriate assignment can come from word frequencies, word-term correlations, a hand crafted rule base (Hayes & Weinstein 1991), a case base (Goodman 1991) or a machine learning algorithm (Chen 1995). In addition to the presence of key words, our approach makes use of the language syntax and semantic scope to help identify concepts in text.

The second dimension concerns the stability of the categories to be assigned. Most approaches assume the categories are pre-defined, however, statistical and vector space methods that create clusters of documents according to some metric (as opposed to an intuitive breakdown), may not retain the same set of categories (clusters) as new documents are added. The principal advantage of clustering is that completely automated classification can be accomplished with no human intervention required. However, the categories (represented by clusters) may not be meaningful to a user. By contrast, our approach retains the classification users are familiar with and provides a more expressive set of classification criteria, but requires a model of the domain be constructed prior to training.

In (Riloff & Lehnert 1994), the authors discuss three information extraction based algorithms for performing a binary classification (relevance or irrelevance to a given domain) of news stories. These approaches use syntactic constraints in the form of a case base to filter the articles. The constraints are generated from a training corpus using a partially automated process that requires manual inspection to determine which constraints are best at discriminating relevant and irrelevant articles.

Our approach extends the work in (Riloff & Lehnert 1994) in several dimensions. First, our approach affords multiple category classification by taking into account the semantic classification of terms in the documents. Second, by building a tree of information extraction based classifiers, we are able to cover a broader domain in an incremental fashion. Third, the decisions (relevant or irrelevant) of each classifier serve to efficiently guide a document to the appropriate class or classes in the leaf nodes.

Conclusions and Future Work

In this paper, we have described a novel multiple category document classification approach that uses Information Extraction techniques to assign novel documents to a set of classes based on concepts present in the text. Both semantic and syntactic constraints on words in the text define concepts to determine the relevancy of a document in a given category.

Our current work involves expanding the prototype system to cover more of the GLIN domain. In a related area of research, we are capitalizing on the resulting classification by creating a conceptual index of the classified documents. Users will then be able to query the indexed collection of documents using concepts to guide their search as opposed to simply keywords.

Acknowledgments

This work has been supported by a grant to Nabil Adam from NASA CESDIS.

The MARMOT, BADGER and CRYSTAL software was developed at the University of Massachusetts Center for Intelligent Information Retrieval (CIIR) and is Copyright 1994, 1995, 1996 ACSIOM.

References

Adam, N. R.; Edelson, B.; El-Ghazawi, T.; Halem, M.; Kalpakis, K.; Kozura, N.; Medina, R.; and Yesha, Y. 1996. Global Legal Information Network (GLIN). *American University Law Review.*

Chen, H. 1995. Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning and Genetic Algorithms. *Journal of the American Scoiety for Information Science* 46(3):194 – 216.

Goodman, M. 1991. Prism: A Case-Based Telex Classifier. In *Proceedings of the Second Annual Conference on Innovative Applications of Artificial Intelligence*, 25 – 37. AAAI Press, MIT Press.

Hayes, P. J., and Weinstein, S. P. 1991. Construe/TIS: A System for Content-Based Indexing of a Database of News Stories. In *Proceedings of the Second Annual Conference on Innovative Applications of Artificial Intelligence*, 49–64. AAAI Press.

Holowczak, R. 1997. *Extractors for Digital Library Objects.* Ph.D. Dissertation, Rutgers University, Department of MS/CIS.

McCarthy, J. 1993. Tools and Techniques for Rapid Prototyping. In *Proceedings of the Fifth Message Understanding Conference*, 347–348. Morgan Kaufman. San Francisco, CA.

Riloff, E., and Lehnert, W. 1994. Information Extraction as a Basis for High-Precision Text Classification. *ACM Transactions on Information Systems* 12(3):296 – 333.

Riloff, E., and Shoen, J. 1995. Automatically Acquiring Conceptual Patterns Without an Annotated Corpus. In *Proceedings of the Third Workshop on Very Large Corpora*, 148 – 161.

Salton, G. 1989. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer.* Reading, MA.: Addison-Wesley Publishing Co.

Soderland, S.; Fisher, D.; Aseltine, J.; and Lehnert, W. 1995. CRYSTAL: Inducing a Conceptual Dictionary. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1314 – 1319.

Uschold, M., and King, M. 1995. Towards a Methodology for Building Ontologies. In *Proceedings of Workshop on Basic Ontological Issues in Knowledge Sharing - In conjunction with the 14th International Joint Conference on Artificial Intelligence 1995.*