Automatic Identification of Rhetorical Roles using Conditional Random Fields for Legal Document Summarization

M. Saravanan
Ph. D Research Scholar
Department of CS & E
IIT Madras, Chennai-36
msdess@yahoo.com

Dr. B. Ravindran
Assistant Professor
Department of CS & E
IIT Madras, Chennai-36
ravi@cse.iitm.ac.in

Dr. S. Raman
Professor (retd.)
Department of CS & E
IIT Madras, Chennai-36
ramansubra@gmail.com

Abstract

In this paper, we propose a machine learning approach to rhetorical role identification from legal documents. In our approach, we annotate roles in sample documents with the help of legal experts and take them as training data. Conditional random field model has been trained with the data to perform rhetorical role identification with reinforcement of rich feature sets. The understanding of structure of a legal document and the application of mathematical model can brings out an effective summary in the final stage. Other important new findings in this work include that the training of a model for one sub-domain can be extended to another sub-domains with very limited augmentation of feature sets. Moreover, we can significantly improve extraction-based summarization results by modifying the ranking of sentences with the importance of specific roles.

1 Introduction

With the availability of large number of colossal legal documents in electronic format, there is a rising need for effective information retrieval tools to assist in organizing, processing and retrieving this information and presenting them in a suitable user-friendly format. To that end, text summarization is an important step for many of these larger information management goals. In recent years, much attention has been focused on the problem of understanding the structure and textual units in legal judgments (Farzindar & Lapalme, 2004). In this case, performing automatic segmentation of a document to understand the rhetorical roles turns out to be an important research issue. For instance, Farzindar (2004) proposed a text summarization method to manipulate factual and heuristic knowledge from legal documents. Hachey and Grover (2005) explored machine learning approach to rhetorical status classification by performing fact extraction and sentence extraction for automatic summarization of texts in the legal domain. They formalized the problem to extract most important units based on the identification of thematic structure of the document and determination of argumentative roles of the textual units in the judgment. They mainly used linguistic features to identify the thematic structures.

In this paper, we discuss methods for automatic identification of rhetorical roles in legal judgments based on rules and on machine learning techniques. Using manually annotated sample documents on three different legal sub-domains (rent control, income tax and sales tax), we train an undirected graphical model to segment the documents along different rhetorical structures. To represent the documents for this work, we mainly used features like cue words, state transition, named entity, position and other local and global features. The segmented texts with identified roles play a crucial part in re-ordering the ranking in the final extraction-based summary. The important sentences are extracted based on the term distribution model given in [Saravanan *et al*, 2006]. In order to develop a generic approach to perform segmentation, we use a fixed set of seven rhetorical categories based on Bhatia's (1993) genre analysis shown in Table 1.

Graphical Models are nowadays used in many text processing applications; however the main

Rhetorical Roles	*Description*
Identifying the case	The sentences that are present in a judgment to identify the issues to be decided for a case. Courts call them as "Framing the issues".
Establishing facts of the case	The facts that are relevant to the present proceedings/litigations that stand proved, disproved or unproved for proper applications of correct legal principle/law.
Arguing the case	Application of legal principle/law advocated by contending parties to a given set of proved facts.
History of the case	Chronology of events with factual details that led to the present case between parties named therein before the court on which the judgment is delivered.
Arguments (Analysis)	The court discussion on the law that is applicable to the set of proved facts by weighing the arguments of contending parties with reference to the statute and precedents that are available.
Ratio decidendi *(Ratio of the decision)*	Applying the correct law to a set of facts is the duty of any court. The reason given for application of any legal principle/law to decide a case is called Ratio decidendi in legal parlance. It can also be described as the central generic reference of text.
Final decision *(Disposal)*	It is an ultimate decision or conclusion of the court following as a natural or logical outcome of ratio of the decision

Table 1. The current working version of the rhetorical annotation scheme for legal judgments.

focus has been performing Natural Language processing tasks on newspaper and research paper domains. As a novel approach, we have tried and implemented the CRF model for role identification in legal domain. In this regard, we have first implemented rule based approach and extend this method with additional features and a probabilistic model. In our work, we are in the process of developing a fully automatic summarization system for a legal domain on the basis of Lafferty's (2001) segmentation task and Teufel & Moen's (2004) gold standard approaches. Legal judgments are different in characteristics compared with articles reporting scientific research papers and other simple domains related to the identification of basic structures of a document. To perform a summarization methodology and find out important portions of a legal document is a complex problem (Moens, 2004). Even the skilled lawyers are facing difficulty in identifying the main decision part of a law report. The genre structure identified for legal judgment in our work plays a crucial role in identifying the main decision part in the way of breaking the document in anaphoric chains. The sentence extraction task forms part of an automatic summarization system in the legal domain. The main focus of this paper is information extraction task based on the identified roles and methods of structuring summaries which has considered being a hot research topic (Yeh *et al*., 2005). Now we will discuss the importance of identifying rules in the data collection by various methods available for rule learning in the next section.

2 Text Segmentation Algorithms

We explain two approaches to text segmentation for identifying the rhetorical roles in legal judgments. The focus of the first approach is on a rule-based method with novel rule sets which we fine-tuned for legal domains. That is, we frame text segmentation as a rule learning problem. The proposed rule-based method can be enhanced with additional features and a probabilistic model. An undirected graphical model, Conditional Random Field (CRF) is used for this purpose. It shows significant improvement over the rule-based method. The explanation of these methods is given in the following sections.

2.1 Rule-based learning algorithms

Most traditional rule learning algorithms are based on a divide-and-conquer strategy. SLIPPER [Cohen, 1999] is one of the standard rule learning algorithms used for information retrieval task. In SLIPPER, the ad hoc metrics used to guide the growing and pruning of rules are replaced with metrics based on the formal analysis of boosting algorithms. For each instance, we need to check each and every rule in the rule set for a given sentence. It takes more time for larger corpora compared to other rule learning algorithms even for a two-class problem. If we need to consider more than two classes and to avoid overfitting of ensemble of rules, one has to think of grouping the rules in a rule set and some chaining mechanism has to be followed. Another rule learning algorithm

RuleFit (Friedman & Popescu, 2005) generates a small comprehensible rule set which is used in ensemble learning with larger margin. In this case, overfitting may happen, if the rule set gets too large and thus some form of control has to be maintained. Our main idea is to find a preferably small set of rules with high predictive accuracy and with marginal execution time.

We propose an alternative rule learning strategy that concentrates on classification of rules and chaining relation in each rhetorical role (Table 1) based on the human annotation schemes. A chain relation is a technique used to identify co-occurrences of roles in legal judgments. In our approach, rules are conjunctions of primitive conditions. As used by the boosting algorithms, a rule set R can be any hypothesis that partitions the set of instance X into particular role categorization; the set of instances which satisfy any one of seven different set of categorized roles. We start by generating rules that describe the original features found in the training set. Each rule outputs 1 if its condition is met, 0 if it is not met. Let us now define for a sample document X = (S1, S2,….,Sm) of size m, we assume that the set of rules R = $\{r_1,r_2,\ldots\}$ are applied to sample X, where each rule $r_i : X \rightarrow L$ represents the mapping of sentences of X onto a rhetorical role and L = $\{L_1,L_2,\ldots,L_7\}$. Each L_i represents a rhetorical role from the fixed set shown in Table 1. An outline of our method is given below.

```
Procedure Test (X)
 {   Read test set
     Read instances from sample X (instances may be
      words, N-grams or even full sentences)
      Apply rules in R (with role categorization
             by maintaining chain relation)
      For k = 1 to m sentences
        For i = 1, 2, …. no. of instances in each sentence
        For j = 1 to 7      /* 7 identified roles */
        If there exist a rule which satisfies then
             X(i,j) gets a value 1
        Else
          X(i,j) gets a value {1,0} based on chain relation
        S(k) = L (argmax Σ(X(i,j)))
                    j   i
}
```

2.2 Conditional Random Fields and Features

The CRF model-based retrieval system designed in this paper will depict the way a human can summarize a legal judgment by understanding the importance of roles and related contents. Conditional Random Fields is one of the recently emerging graphical models which have been used for text segmentation problems and proved to be one of the best available frame works compared to other existing models (Lafferty, 2001). A judgment can be regarded as a sequence of sentences that can be segmented along the seven rhetorical roles where each segments is relatively coherent in content. We use CRF as a tool to model the text segmentation problem. CRFs are undirected graphical models used to specify the conditional probabilities of possible label sequences given an observation sequence. Moreover, the conditional probabilities of label sequences can depend on arbitrary, non independent features of the observation sequence, since we are not forming the model to consider the distribution of those dependencies. In a special case in which the output nodes of the graphical model are linked by edges in a linear chain, CRFs make a first-order Markov independence assumption with binary feature functions, and thus can be understood as conditionally-trained finite state machines (FSMs) which are suitable for sequence labeling.

A linear chain CRF with parameters C = $\{C_1,C_2,\ldots..\}$ defines a conditional probability for a label sequence l = $l_1,\ldots..l_W$ (e.g., Establishing facts of the case, Final decision, etc.) given an observed input sequence s = $s_1,\ldots s_W$ to be

$$P_C(l \mid s) = \frac{1}{Z_s} \exp\left[\sum_{t=1}^{w} \sum_{k=1}^{m} C_k f_k(l_{t-1}, l_t, s, t)\right] \quad \ldots. \quad (1)$$

where Z_s is the normalization factor that makes the probability of all state sequences sum to one, $f_k(l_{t-1}, l_t, s, t)$ is one of m feature functions which is generally binary valued and C_k is a learned weight associated with feature function. For example, a feature may have the value of 0 in most cases, but given the text "points for consideration", it has the value 1 along the transition where l_{t-1} corresponds to a state with the label *Identifying the case*, l_t corresponds to a state with the label *History of the case*, and f_k is the feature function PHRASE= "points for consideration" belongs to s at position t in the sequence. Large positive values for C_k indicate a preference for such an event, while large negative values make the event unlikely and near zero for relatively uninformative features. These weights are set to maximize the conditional log likelihood of labeled sequence in a training set D = $\{(s_t, l_t) : t = 1,2,\ldots w)$, written as:

$$L_C(D) = \sum_i \log P_C(l_i \mid s_i)$$

$$= \sum_i \left(\sum_{t=1}^{w} \sum_{k=1}^{m} C_k f_k(l_{t-1}, l_t, s, t) - \log Z s_i \right) \ldots (2)$$

The training state sequences are fully labeled and definite, the objective function is convex, and thus the model is guaranteed to find the optimal weight settings in terms of L_C (D). The probable labeling sequence for an input s_i can be efficiently calculated by dynamic programming using modified Viterbi algorithm. These implementations of CRFs are done using newly developed java classes which also use a quasi-Newton method called L-BFGS to find these feature weights efficiently. In addition to the following standard set of features, we also added other related features to reduce the complexity of legal domain.

Indicator/cue phrases – The term 'cue phrase' indicates the key phrases frequently used which are the indicators of common rhetorical roles of the sentences (e.g. phrases such as "We agree with court", "Question for consideration is", etc.,). In this study, we encoded this information and generated automatically explicit linguistic features. Feature functions for the rules are set to 1 if they match words/phrases in the input sequence exactly.

Named entity recognition - This type of recognition is not considered fully in summarizing scientific articles (Teufel & Moens, 2002). But in our work, we included few named entities like Supreme Court, Lower court etc., and generate binary-valued entity type features which take the value 0 or 1 indicating the presence or absence of a particular entity type in the sentences.

Local features and Layout features - One of the main advantages of CRFs is that they easily afford the use of arbitrary features of the input. One can encode abbreviated features; layout features such as position of paragraph beginning, as well as the sentences appearing with quotes, all in one framework.

State Transition features - In CRFs, state transitions are also represented as features (Peng & McCullam, 2006). The feature function $f_k(l_{t-1}, l_t, s, t)$ in Eq. (1) is a general function over states and observations. Different state transition features can be defined to form different Markov-order structures. We define state transition features corresponding to appearance of years attached with Section and Act nos. related to the labels *Arguing the case* and *Arguments*.

Legal vocabulary features - One of the simplest and most obvious set of features is decided using the basic vocabularies from a training data. The words that appear with capitalizations, affixes, and in abbreviated texts are considered as important features. Some of the phrases that include **v.** and *act/section* are the salient features for *Arguing the case* and *Arguments* categories.

2.3 Experiments with role identification

We have gathered a corpus of legal judgments up to the year 2006 which were downloaded from www.keralawyer.com specific to the sub-domains of rent control, income tax and sales tax. Using the manually annotated subset of the corpus (200 judgments) we have performed a number of preliminary experiments to determine which method would be appropriate for role identification. The annotated corpus is available from iil.cs.iitm.ernet.in/datasets. Even though, income tax and sales tax judgments are based on similar facts, the number of relevant legal sections / provisions are differ. The details and structure of judgments related to rent control domain are not the same compared to income tax and sales tax domains. Moreover, the roles like ratio decidendi and final decision occur many times spread over the full judgment in sales tax domain, which is comparatively different to other sub-domains. We have implemented both the approaches on rent control domain successfully. We found that the other sub-domains need specific add-on features which improve the result by an additional 20%. Based on this, we have introduced additional features and new set of rules for the income tax and sales tax related judgments. The modified rule set and additional features are smaller in number, but create a good impact on the rhetorical status classification in the sales tax and income tax domains. It is common practice to consider human performances as an upper bound for most of the IR tasks, so in our evaluation, the performance of the system has been successfully tested by matching with human annotated documents.

Kappa (Siegal & Castellan, 1988) is an evaluation measure used in our work to compare the inter-agreement between sentences extracted by two

	Rhetorical Roles	Precision			Recall			F-measure		
		Slipper	Rule-based	CRF	Slipper	Rule-based	CRF	Slipper	Rule-based	CRF
Rent Control Domain	Identifying the case	0.641	0.742	0.846	0.512	0.703	0.768	0.569	0.722	0.853
	Establishing the facts of the case	0.562	0.737	0.824	0.456	0.664	0.786	0.503	0.699	0.824
	Arguing the case	0.436	0.654	0.824	0.408	0.654	0.786	0.422	0.654	0.805
	History of the case	0.841	0.768	0.838	0.594	0.716	0.793	0.696	0.741	0.815
	Arguments	0.543	0.692	0.760	0.313	0.702	0.816	0.397	0.697	0.787
	Ratio of decidendi	0.574	0.821	0.874	0.480	0.857	0.903	0.523	0.839	0.888
	Final Decision	0.700	0.896	0.986	0.594	0.927	0.961	0.643	0.911	0.973
							Micro-Average of F-measure	**0.536**	**0.752**	**0.849**
	Rhetorical Roles	Precision			Recall			F-measure		
		Slipper	Rule-based	CRF	Slipper	Rule-based	CRF	Slipper	Rule-based	CRF
Income Tax Domain	Identifying the case	0.590	0.726	0.912	0.431	0.690	0.852	0.498	0.708	0.881
	Establishing the facts of the case	0.597	0.711	0.864	0.512	0.659	0.813	0.551	0.684	0.838
	Arguing the case	0.614	0.658	0.784	0.551	0.616	0.682	0.581	0.636	0.729
	History of the case	0.437	0.729	0.812	0.418	0.724	0.762	0.427	0.726	0.786
	Arguments	0.740	0.638	0.736	0.216	0.599	0.718	0.334	0.618	0.727
	Ratio of decidendi	0.416	0.708	0.906	0.339	0.663	0.878	0.374	0.685	0.892
	Final Decision	0.382	0.752	0.938	0.375	0.733	0.802	0.378	0.742	0.865
							Micro-Average of F-measure	**0.449**	**0.686**	**0.817**
	Rhetorical Roles	Precision			Recall			F-measure		
		Slipper	Rule-based	CRF	Slipper	Rule-based	CRF	Slipper	Rule-based	CRF
Sales Tax Domain	Identifying the case	0.539	0.675	0.842	0.398	0.610	0.782	0.458	0.641	0.811
	Establishing the facts of the case	0.416	0.635	0.784	0.319	0.559	0.753	0.361	0.595	0.768
	Arguing the case	0.476	0.718	0.821	0.343	0.636	0.747	0.399	0.675	0.782
	History of the case	0.624	0.788	0.867	0.412	0.684	0.782	0.496	0.732	0.822
	Arguments	0.500	0.638	0.736	0.438	0.614	0.692	0.467	0.626	0.713
	Ratio of decidendi	0.456	0.646	0.792	0.318	0.553	0.828	0.375	0.596	0.810
	Final Decision	0.300	0.614	0.818	0.281	0.582	0.786	0.290	0.598	0.802
							Micro-Average of F-measure	**0.407**	**0.637**	**0.787**

Table 2. Precision, Recall and F-measure for seven rhetorical roles

human annotators for role identification in legal judgments. The value (K=0.803) shows the good reliability of human annotated corpus. The results given in Table 2 show that CRF-based and rule-based methods perform well for each role categories compared to SLIPPER method. CRF-based method performs extremely well and paired t-test result indicates that it is significantly ($p < .01$) higher than the other two methods on rhetorical role identification for legal judgments belonging to rent control, income tax and sales tax sub-domains.

3 Legal Document Summarization

Extraction of sentences in the generation of a summary at different percentage levels of text is one of the widely used methods in document summarization (Radev *et al.*, 2002). For the legal domain, generating a summary from the original judgment is a complex problem. Our approach to produce the summary is extraction-based method which identifies important elements present in a legal judgment. The identification of the document structure using CRF-model categorizes the key ideas from the details of a legal judgment. The genre structure has been applied to final summary to improve the readability and coherence. In order to evaluate the effectiveness of our summarizer, we have applied four different measures to look for a match on the model summary generated by humans (*head notes*) from the text of the original judgments.

3.1 Applying term distribution model

The automatic text summarization process starts with sending legal document to a preprocessing stage. In this preprocessing stage, the document is to be divided into segments, sentences and tokens. We have introduced some new feature identification techniques to explore paragraph alignments. This process includes the understanding of abbreviated texts and section numbers and arguments

which are very specific to the structure of legal documents. The other useful statistical natural language processing tools, such as filtering out stop list words, stemming etc., are carried out in the preprocessing stage. The resulting intelligible words are useful in the normalization of terms in the term distribution model (Saravanan *et al.*, 2006). During the final stage, we have altered the ranks or removed some of the sentences from the final summary based on the structure discovered using CRF. The summarization module architecture is shown in Figure 1.

The application of term distribution model brings out a good extract of sentences present in a legal document to generate a summary. The sentences with labels identified during CRF implementation can be used with the term distribution model to give more significance to some of the sentences with specific roles. Moreover, the structure details available in this stage are useful in improving the coherency and readability among the sentences present in the summary.

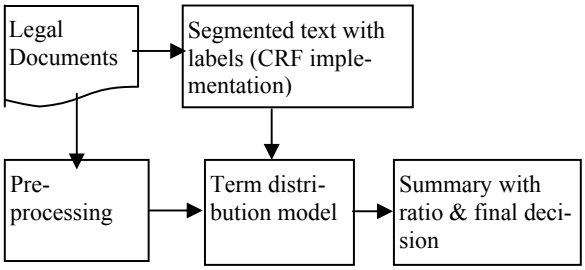

Figure 1.Architectural view of summarization system.

3.2 Evaluation of a summary

Extrinsic and intrinsic are the two different evaluation strategies available for text summarization (Sparck Jones & Gablier, 1996). Intrinsic measure shows the presence of source contents in the summary. F-measure and MAP are two standard intrinsic measures used for the evaluation of our system-generated summary. We have also used ROUGE evaluation approach (Lin, 2004) which is based on n-gram co-occurrences between machine summaries and *ideal* human summaries. In this paper, we have applied ROUGE-1 and ROUGE-2 which are simple n-gram measures. We compared our results with Microsoft, Mead Summarizer (Radev et al., 2003) and other two simple baselines: one which chooses 15% of words of the beginning of the judgment and second chooses last ginning of the judgment and second chooses last 10% of words of the judgment with human reference summaries. Both the baselines defined in this study are standard baselines for newspaper and research domains. The result shown in Table 3 highlights the better performances of our summarizer compared to other methods considered in this study. We can see that the results of MEAD and WORD summaries are not at the expected level, while our summarizer is best in terms of all four evaluation measures. Results are clearly indicated that our system performs significantly better than the other systems for legal judgments.

	MAP	F-measure	ROUGE-1	ROUGE-2
Baseline 1	0.370	0.426	0.522	0.286
Baseline 2	0.452	0.415	0.402	0.213
Microsoft Word	0.294	0.309	0.347	0.201
Mead	0.518	0.494	0.491	0.263
Our system	0.646	0.654	0.685	0.418

Table 3. MAP, F-measure and ROUGE scores.

4 Conclusion

This paper describes a novel method for generating a summary for legal judgments with the help of undirected graphical models. We observed that rhetorical role identification from legal documents is one of the primary tasks to understand the structure of the judgments. CRF model performs much better than rule based and other rule learning method in segmenting the text for legal domains. Our approach to summary extraction is based on the extended version of term weighting method. With the identified roles, the important sentences generated in the probabilistic model will be reordered or suppressed in the final summary. The evaluation results show that the summary generated by our summarizer is closer to the human generated head notes, compared to the other methods considered in this study. Hence the legal community will get a better insight without reading a full judgment. Moreover, our system-generated summary is more useful for lawyers to prepare the case history related to presently appearing cases.

References

Atefeh Farzindar and Guy Lapalme. 2004. *Legal text summarization by exploration of the thematic structures and argumentative roles,* In Text summarization Branches out workshop held in conjunction with ACL 2004, pages 27-34, Barcelona, Spain.

Atefeh Farzindar and Guy Lapalme. 2004. *Letsum, an automatic legal text summarizing system*, Legal Knowledge and Information System, Jurix 2004: The Seventeenth Annual Conference, Amsterdam, IOS Press, PP.11-18.

Ben Hachey and Claire Grover. 2005. *Sequence Modeling for sentence classification in a legal summarization system*, Proceedings of the 2005 ACM symposium on Applied Computing.

Saravanan , M., Ravindran, B. and Raman, S. 2006. *A Probabilistic Approach to Multi-document summarization for generating a Tiled Sumamry*, International Journal of Computational Intelligence and Applications, 6(2): 231-243, Imperial College Press.

Bhatia, V.K., 1999. *Analyzing Genre: Language Use in Professional Settings*, London, Longman.

John Lafferty, Andrew McCullam and Fernando Pereira, 2001. *Conditional Random Fields: Probabilistic models and for segmenting and labeling sequence data*, Proceedings of international conference on Machine learning.

Simone Teufel and Marc Moens, 2002. *Summarizing scientific articles – experiments with relevance and rhetorical status,* Association of Computational Linguistics, 28(4): 409-445.

Marie-Francine Moens, 2004. *An Evaluation Forum for Legal Information Retrieval Systems*? Proceedings of the ICAIL-2003 Workshop on Evaluation of Legal Reasoning and Problem-Solving Systems (pp. 18-24). International Organization for Artificial Intelligence and Law.

Yen-Yuan Yeh, Hao-Ren Ke, Wei-Pang Yang, and I-Heng Meng, 2005. *Text summarization using a trainable summarizer and latent semantic analysis*, Information processing management, 41(1):75-95.

Cohen,W., and Singer, Y. 1999. *A simple, fast, and effective rule learner*, Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99), AAAI Press, pp.335-342.

Friedmen, J.H., & and Popescu, B. E. 2005. *Predictive learning via rule ensembles* (Technical Report), Stanford University.

Fuchun Peng and Andrew McCullam, 2006. *Accurate information extraction from research papers using conditional random fields*, Information Processing Management, 42(4): 963-979.

Saravanan , M., Ravindran, B. and Raman, S. 2006. *Improving legal document Summarization using graphical models*, Legal Knowledge and Information System, JURIX 2006: The Nineteenth Annual Conference, Paris, IOS Press, PP.51-60.

Siegal, Sidney and N.John Jr. Castellan. 1988. *Nonparametric statistics for the behavioral sciences*, McGraw Hill, Berkeley, CA.

Dragomir Radev, Eduard Hovy, Kathleen McKeown. 2002. *Introduction to the special issue on summarization,* Computational Linguistics 28(4)4, Association for Computing Machinery.

Karen Sparck Jones and Julia Galliers. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Natural Language Engineering, 4(2):175–190, Springer-Verlag.

Lin, Chin-Yew. 2004. *ROUGE: a Package for Automatic Evaluation of Summaries,* Proceedings of Workshop on Text Summarization, pp: 21--26, Barcelona, Spain.

Dragomir Radev, Jahna Otterbaher, Hong Qi, and Daniel Tam. 2003. Mead Reducs: Michigan at duc 2003. In DUC03, Edmonton, Alberta, Canada, May 31- June 1. Association for Computational Linguisics.