Automated Methods for Extracting and Expanding Lists in Regulatory Text

Alan BUABUCHACHART, Nina CHARNESS,
Katherine METCALF, and Leora MORGENSTERN

Leidos, 4001 N. Fairfax Drive, Arlington, VA 22203

Abstract. This is a highly condensed version of a paper [1] that presents automated methods to accurately transform regulations with bulleted lists into sets of complete sentences that include their proper context. We discuss the technical challenges addressed, including extracting intended structure from HTML documents, and correctly distributing preambles over nested text. Our work has been used to preprocess the corpus used for our experiments in classifying paragraphs in regulatory documents by several categories, including illocutionary point, regulation type, and reference structure. That work is presented in a companion paper published in the JURIX 2013 proceedings.

Keywords. text analysis, regulation, preprocessing, classification

1. Introduction

Many regulatory documents contain or are largely composed of bulleted lists. Such lists break up complex text and increase comprehension for human readers, who are good at distributing preambles over bulleted text as they read. However, automated processing of such documents is difficult, especially if bulleted units are not complete sentences, and if there are multiple levels of nesting. We develop automated methods to transform text that contains bullets to text in which the bulleted text is fully expanded, with preambles distributed over the bulleted text.

This work is a preliminary step in our study of the feasibility of automating the translation of regulatory text into formal, executable rules. Our approach to this general problem involves both machine learning and deep parsing techniques; we have found that distribution is a necessary first step for both tasks. As discussed in [2], both the consistency of annotation/training data and the performance of clustering algorithms is superior when using expanded text rather than standard bulleted text, or bulleted text to which sentence splitting techniques [3] have been applied. Moreover, the loss of context inherent in sentence splitting suggests that expanded text will lead to more accurate parsing.

2. Motivating Example: The importance of context in reading bulleted lists

Domain and corpus: We are working with a corpus of 250 United States financial regulation units. Consider, e.g., the initial fragment of FINRA Rule 3240:

```
3240. Borrowing From or Lending to Customers
(a) Permissible Lending Arrangements; Conditions
 No person associated with a member in any registered capacity may borrow money from or lend money
 to any customer of such person unless:
    (1) the member has written procedures allowing the borrowing and lending of money between
such registered persons and customers of the member;
    (2) the borrowing or lending arrangement meets one of the following conditions:
      (A) the customer is a member of such person's immediate family;
      (B) the customer (i) is a financial institution regularly engaged in the business of
       providing credit ...
       and (ii) is acting in the course of such business;
      (C) the customer and the registered person are both registered persons of the same member; .....
```

Bulleted structure aids human comprehension by breaking up text. We understand that bullet (a) lists ways in which lending is allowed; that subbullet (2) specifies alternative necessary conditions constraining the relationship between customer and lender. As we read the text we must keep *context* in mind.

[3] and [4] advocate processing bulleted text by using punctuation cues to do sentence splitting. This yields sentences such as *the member has written procedures allowing the borrowing and lending of money between such registered persons and customers of the member*. Such sentences are missing context and are therefore difficult to understand.

3. Extracting from HTML, Tree Building, Distributing Preambles

In developing our technical approach, we address two hard problems: (1) extracting bulleted structure from available text; (2) building a tree structure that supports expansion and distribution of parent preambles over child bullets for arbitrarily deep levels of nesting. We can then traverse the tree to obtain the distributed text.

We recovered bulleted structure using HTML files from 6 different online law sources. Utilities like jsoup facilitate detection of paragraphs and indentation. HTML tags facilitate getting rid of junk text. Unfortunately, no source HTML files use standard bulleting tags (e.g., <ol>,) to indicate bullets in the text. Recovering the bulleted list structure is challenging because each website has its own conventions for representing lists, necessitating customized analysis. One source often has several nested labels appearing in a single line, which makes it difficult to distinguish bullet labels from references to other regulation parts and introduces potential error. For all sources, it is difficult to determine if a label like "(i)" acts as a letter or a Roman numeral, which could introduce error when multiple levels of nesting are present.

The extraction step outputs a set of labels, each of which is assigned a label type (e.g., uppercase letter, Arabic numeral) and is attached to a chunk of text in the document. The tree is then built by traversing the document:

For each paragraph

If the label type is different than the previous label type

If the label type is not on the stack
Create a new node and add it as a child of the previous node
Save previous node as the parent of this node
Put this label type on the stack
Else
Remove everything above this label type from the stack
Find the parent of the current label type
Create a new node and add it as a child of that parent
Else
Create a new node and add it as a child of the same parent of previous node

It is then easy to distribute preambles over bullet content: every path in the tree corresponds to one fully expanded bullet. One need only read out the text associated with the nodes in the path to obtain the fully expanded and distributed bullet. The text associated with all ancestors of the bullet is concatenated with the text of the bullet itself. A sample of the results for the distributed version of our example is shown below.

3240. Borrowing From or Lending to Customers (a) Permissible Lending Arrangements; Conditions No person associated with a member in any registered capacity may borrow money from or lend money to any customer of such person unless: (2) the borrowing or lending arrangement meets one of the following conditions: (B) the customer (i) is a financial institution regularly engaged in the business of providing credit

4. Results and Utility

We have achieved near perfect results in distribution of bulleted text. We have used this method to preprocess our corpus of 250 regulation units, and have found that annotation and clustering algorithms are markedly superior when working on text in which bullets have been expanded [2]). When using sentence splitting methods, we could identify definitions with an average F1 score of barely .8. (Recall was relatively low since many definitions were identified as regulations.) Using the expanded, distributed text, the F1 score rose to .95. Certain classification experiments were impossible before bullet expansion. For example, we could not annotate regulation types after sentence splitting, since the lines of text often had too little context; these difficulties disappeared once bulleted lists were expanded.

5. Acknowledgements

The research described in this paper is supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory (AFRL) contract number FA8750-13-C-0085. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.
Thanks also to Adam Wyner, Ron Keesing, and Ted Senator for helpful ideas and suggestions.

References

[1] A. Buabuchachart, N. Charness, K. Metcalf, and L. Morgenstern, Automated Methods for Extracting and Expanding Lists in Regulatory Text, Working paper, at http://cs.nyu.edu/leora/papers .
[2] A. Buabuchachart, K. Metcalf, N. Charness, and L. Morgenstern, Automated Classification of Regulatory Text by Discourse Structure, Reference Structure, and Regulation Type, *JURIX 2013*.
[3] F. Dell'Orletta, S. Marchi, S. Montemagni, B. Plank, and G. Venturi, The SPLeT-2012 Shared Task on Dependency Parsing of Legal Texts, *SPLeT 2012, Workshop on Semantic Processing of Legal Text*, at *LREC 2012*, Istanbul.
[4] A. Wyner and W. Peters, On Rule Extraction from Regulations, *JURIX 2011*.