NLP for Privacy Policy Document Analysis

Aaron Massey, UMBC

Alden Dima, NIST
NLP in Requirements Engineering

Mid-2010s: NLP is Resurgent

“[W]e witness the novel golden age of NLP technologies [...] It is therefore an appropriate moment to create a venue in which researchers on applications of NLP to RE problems can meet, share ideas and create synergies, assisted by experts from the NLP community.”

F. Dalpiaz, A. Ferrari, X. Franch, and C. Palomares, Call for Papers, 1st Workshop on Natural Language Processing for Requirements Engineering, 2018

Mid-1990s: NLP is Worthless

“[N]atural language does not now, nor will it in the foreseeable future, provide a level of understanding that could be relied upon, and even if it could, it is highly questionable that the resulting system would be of great use in requirements engineering.”

Primary Ongoing Challenge for NLP in SE/RE

“[P]erhaps a dumb tool doing an identifiable part of such a task may be better than an intelligent tool trying but failing in unidentifiable ways to do the entire task.”


Measurement of NLP techniques are misleading for Software Engineering and Requirements Engineering:

**Option 1:** Break the problem down into smaller tasks with a smaller penalty

**Option 2:** Adjust measurement techniques to account for SE/RE

**Option 3:** Focus on Human-in-the-Loop systems
Automated Text Mining for Requirements Analysis of Policy Documents

Goal: Identification of documents relevant to specific requirements

Can automated text mining help requirements engineers determine whether a policy document contains requirements expressed as either privacy protections and vulnerabilities?

Uses a large corpus of privacy policy documents:

The RE 2013 Policy Document Corpus consists of 2,061 Privacy Policies, Terms of Use, Terms and Conditions, and Terms of Service documents, and other policy documents.

Automated Text Mining for Requirements Analysis of Policy Documents (2)

Methodology

1. Created a Latent Dirichlet Allocation (LDA) model with $k = 154$ topics using R topicmodels
   - Split docs: training (90%) and holdout (10%)
   - Generated 35 topic models with varying $k$ and picked one with lowest perplexity
2. Identified documents addressing concerns using goals-based RE

Can limit search from 2,061 documents to less than 100 related to given goal keyword

<table>
<thead>
<tr>
<th>Goal Keywords</th>
<th>Topics with Term</th>
<th>Documents with Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALLOW</td>
<td>20</td>
<td>YouTube Terms of Service,</td>
</tr>
<tr>
<td>COLLECT</td>
<td>68</td>
<td>Microsoft Privacy Statement,</td>
</tr>
<tr>
<td>CUSTOMIZE</td>
<td>150</td>
<td>ConocoPhillips Legal and Privacy Statement</td>
</tr>
<tr>
<td>DISCLOSE</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>INFORM</td>
<td>125</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Policy Documents (out of 2,061) Identified as Potentially Containing Goal Statements</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Keyword</strong></td>
</tr>
<tr>
<td>access</td>
</tr>
<tr>
<td>collect</td>
</tr>
<tr>
<td>display</td>
</tr>
<tr>
<td>inform</td>
</tr>
<tr>
<td>opt-in</td>
</tr>
<tr>
<td>request</td>
</tr>
<tr>
<td>specify</td>
</tr>
</tbody>
</table>
Analysis of Privacy Concerns Across Industry

Researchers and regulators: Can we identify privacy concerns within policy documents at scale?

Manual analysis of privacy policy collections
- Example: Professor and eight law students reviewed 249 policies (Marotta-Wurgler, 2016)

Is automated replication possible?
- RE 2013 Policy Document Corpus has 2061 documents (Massey et al, 2013)
- Web crawling enables large collections
  - 1M policy documents

Challenges

Keyword Search → False Positives
Want to quickly re-analyze
- New concerns
- New examples

Classifiers
- Imbalanced classes
- Many concerns → many classifiers
- Lack of labelled training data
  - Often small sets of examples
  - Deep Learning → Large training sets + time
Methodology

1. Extract sentences from training data
2. Create Latent Semantic Indexing (LSI) model of sentences
3. Filter training data using keywords
4. Summarize filtered data with LexRank
5. Manually select exemplar (seed) sentences from filtered data
6. Search of new corpus using exemplar sentences and LSI model
7. Select sentences above cosine similarity threshold
Summary

NLP can play an important role in Requirements Engineering, but the benefits and challenges of this approach remain mostly a topic for academic research.
Thank you!

Aaron Massey
Assistant Professor of Software Engineering
Information Systems Department
University of Maryland Baltimore County
akmassey@umbc.edu
https://userpages.umbc.edu/~akmassey/