Semantic Analysis of Contractual Agreements to Support End-User Interpretation

Najmeh Mousavi Nejad

Smart Data Analytics (SDA) - University of Bonn and
Fraunhofer Intelligent Analysis and Information Systems (IAIS), Germany
nejad@cs.uni-bonn.de

Abstract. The ubiquitous availability of the Internet results in a huge number of apps, software and online services with accompanying contractual agreements in the form of 'terms of use' and 'privacy policy'. Although everyone is exposed to such consent forms, the majority tend to ignore them due to their length and complexity. In this thesis, we focus on interpretation of contractual agreements for the benefit of end-users. By applying text mining and semantic technologies, we develop an approach that extracts important information and retrieves helpful links and resources for the better comprehension. Our approach is based on ontology-based information extraction and machine learning and delivers the unpleasant consent form in a user friendly and visualized format. The evaluation shows that although semi-automatic approaches lead to information loss, they save time and effort by producing instant results and facilitate the end-users' understanding of legal texts.

Keywords: Contractual agreement · Terms of use · Privacy policy · Ontology-based information extraction · Machine learning.

1 Problem Statement

The increasing availability of online services and mobile apps has led to a huge proliferation of terms and conditions regulating their use. In the digital age everyone is exposed to such terms, and in their majority this constitutes ordinary people with limited to no knowledge of legal terms. The problem arises when people ignore the consent forms due to their length and complex terminology. In a recent study, "The biggest lie on the Internet", 543 students were asked to agree to a privacy policy and terms of use in order to join a fictitious social network [6]. Although 26% did not choose the 'quick join', the average time of reading was only 73 seconds. Ignoring these terms is a risk, taken by most users. According to *Skandia*[1], 10% were bound by a longer contract than they expected and 5% lost money by not being able to cancel or amend their bookings.

In order to facilitate the process of digesting terms and conditions for regular end-users, we consider applying text mining and use of domain ontologies

[1] http://www.prnewswire.co.uk/news-releases/skandia-takes-the-terminal-out-of-terms-and-conditions-145280565.html

to provide visualized summaries. The approach considered is broadly applicable to other forms of text-based contractual agreements. However, in this thesis we specifically focus on terms of use (aka. End-User License Agreement or EULA) and privacy policies, since they have the broadest impact and affect everyone. Our research questions which will be answered in the sequel, specifically include: **1) Dose text mining techniques for extracting and summarizing important information from consent forms lead to information loss? and 2) Does our approach need less time and effort for contractual agreements comprehension?**

2 State of the Art

Consent forms such as terms of use and privacy policies have clear structure and terminologies. Therefore, OBIE is a fitting method for processing such texts, since the mappings between natural language text and machine-understandable conceptualizations is more straightforward. OBIE uses an ontology to guide the IE pipeline and annotates the text with the ontology concepts. In recent years, along with the increasing emergence of domain ontologies, OBIE has gained a lot of interest. According to a survey of OBIE applications [8], the most widely-used tools for OBIE are GATE[2], sProUT[3] and the Stanford CoreNLP[4]. We have chosen GATE due to its excellent support for OBIE.

Our literature review covers specifically license agreements and privacy policy studies. `Tl;drLegal`[5] is an online service that uses a manual and crowdsourced way to present a summary of popular EULAs. Furthermore, NLL2RDF is a first attempt which employs NLP and ML techniques to generate RDF expressions of license agreements [1]. The framework is evaluated against a goldstandard which was created manually using Open Digital Right Language (ODRL)[6] and CC REL[7] ontologies. However, NLL2RDF is able to generate only a few number of rights and conditions due to the incomplete training data.

Some efforts have specifically studied privacy policies [2, 4, 5]. A common approach is to use predefined categories and supervised ML to assign classes to policy paragraphs. Furthermore these categories are helpful for assessing the completeness of privacy policies. The primary limitation of these studies is a lack of sufficient training data. The only proper dataset was created by the Usable Privacy Policy Project[8]. OPP-115 contains 115 privacy policies from American companies and was annotated by 3 experts into 10 categories [7]. Polisis[9] exploits OPP-115 to process the privacy policies and presents them in a visualized format.

[2] https://gate.ac.uk/

[3] http://sprout.dfki.de/

[4] https://stanfordnlp.github.io/CoreNLP/

[5] https://tldrlegal.com/

[6] https://www.w3.org/community/odrl/

[7] https://creativecommons.org/ns

[8] https://usableprivacy.org/

[9] https://pribot.org/polisis

To the best of our knowledge, the missing chain in Polisis is analyzing a policy's risk factor and its compliance with the law. In future, we plan to apply ML using OPP-115 to assign a risk score to privacy policies and identify potential mappings between a specific policy and data protection legislation.

3 Proposed Approach

After a thorough literature review covering terms of use (or EULAs) and privacy policies, we considered different approaches to identify the ones that are more suitable for each type of agreement and the in/availability of training corpora. For privacy policy analysis, we rely on the OPP-115 dataset to apply supervised ML and train a reliable model. In contrast, there is no annotated corpus for terms of use, and the creation of one poses a major challenge because they don't follow a specific structure and their scope is generally broader. Depending on the type of an asset (software, website, digital products, etc.), the terms of use differ significantly. They may contain copyright conditions, specific rules on accessing the service, intellectual property rights and various other content. However they all share a common characteristic: they are written using legal terminology — from which it is able to extract a common structure. Based on this assumption, we apply OBIE for extracting pre-defined classes of information.

Having investigated the existing EULA ontologies and vocabularies, we have chosen ODRL as the main ontology for our OBIE pipeline. It is specified in W3C recommendations and has also demonstrated the highest community endorsement. Although the focus of ODLR is digital content, it is broad enough to cover different types of resources. Benefiting from `Permission`, `Prohibition` and `Duty` classes and their properties (e.g., `hasAction`), we define tailored rules utilizing GATE JAPE grammar [3]. Based on repeated observations and consultation with legal experts, we enhanced the ontology to expand the coverage of our rules, e.g., some instances are added to the `Action` class (which is the 'range' of `hasAction` property). Our final framework extracts `Permissions`, `Prohibitions` and `Duties` from an EULA.

Although OBIE is a standard approach, there has been no prior study utilizing OBIE for EULAs. From this point of view our application is new. Furthermore, since we are benefiting from a standard 'model' of the domain, there is a huge potential to better structure similar documents along the same taxonomy. Moreover, having a vocabulary to cover such legal texts can become a standard for structuring also new documents (and not just the existing ones).

4 Methodology

Our framework consists of two separate modules: *EULAide* is responsible for processing license agreements (or terms of use) and is based on OBIE; and *KnIGHT* assigns pre-defined categories to a privacy policy paragraphs and is built upon a supervised ML approach. Figure 1 shows the high-level architecture of our framework and each module is presented in the following subsections.

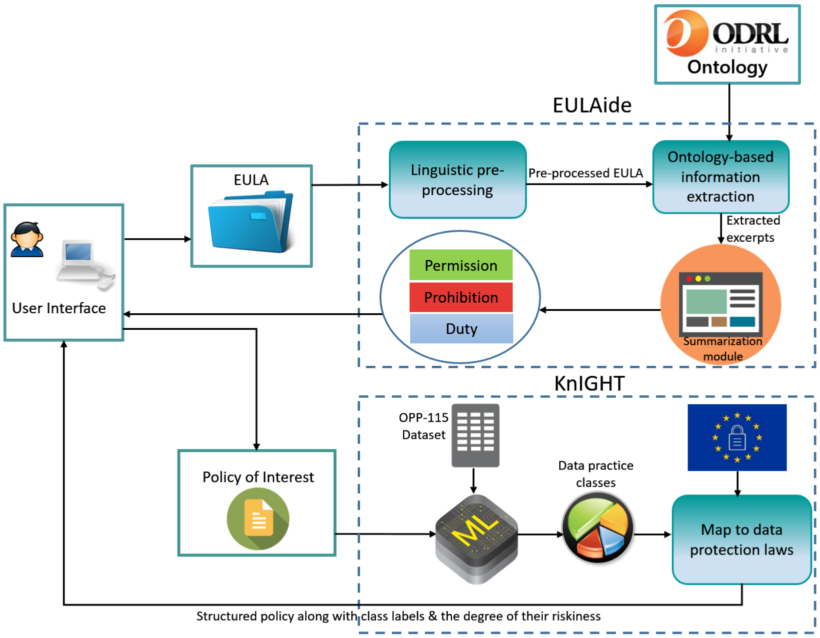


Fig. 1: Architecture and Workflow of Our Approach

4.1 *EULAide*

EULAide extracts important excerpts from EULAs. As shown in the picture, a pre-processing module performs common NLP tasks: tokenizer, sentence splitter, POS tagger, root finder and a text-file gazetteer which contains important keywords from EULAs (e.g., different synonyms for the terms 'license' and 'asset'). The pre-processed EULA will be ingested to the OBIE pipeline, which contains JAPE hand-coded grammar rules based on ODRL community specification documentation[10]. Till now, we have implemented 15 rules, some of which are:

i) *Ontology-based annotation*: separates all ontology-derived *Action* instances into *DutyAction* & *PermProhAction* based on the ontology specification;
ii) *ExtractPermWords*: identifies the important keywords for permissions detection, e.g., *may*, *can*, *grant*, *permit*, etc.;
iii) *ExtractPermission*: extracts the whole sentence, if the pattern is matched.

Table 1 shows the steps towards extracting of a sample `permission`. After the pre-processing phase, first the text-file gazetteer produces two annotations: *License* & *Asset*. Second, the *ontology-based annotation* generates *PermAction* annotation. Third, the *extractPermWords* rule fires and *PermWord* annotation is created. Finally, *extractPermission* detects the whole sentence as a `Permission`.

[10] https://www.w3.org/TR/odrl-vocab/

Table 1: Example of a Permission as extracted by EULAide

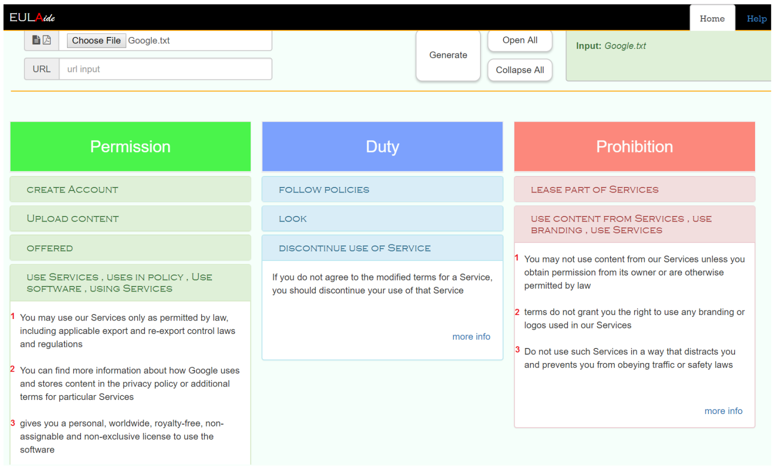
Text-file Gazetteer	**This license** License	grants	you	to copy, share and reproduce	**the product** Asset
Ontology based annotation	This license	grants	you	**to copy, share and reproduce** PermAction	the product
Extract Perm Words	This license	**grants** PermWord	you	to copy, share and reproduce	the product
Extract Permissions	**This license** License	**grants** PermWord	**you** Obj	**to copy, share and reproduce** (PermAction)+	**the product** Asset

The summarization component clusters the similar extracted excerpts and creates a short description for each cluster. Figure 2 shows an example of *EULAide* output. The number of extracted excerpts by OBIE pipeline is 14, whereas the summarization module has reduced the number of clusters to 9.

In order to evaluate the efficiency of *EULAide* we conducted an experiment to identify if the solution enables end-users to invest less time and effort to sufficiently comprehend it. At the same time, we wanted to identify the trade-off between the added support and the information loss expected when applying semi-automatic IE and summarization. As a first step, a corpus containing twenty EULAs in their natural language texts was compiled. Two annotators familiar with EULA texts annotated the corpus independently following an introduction to the relevant ODRL concepts. The Inter-Annotator Agreement (IAA) between two annotators is 90%, which indicates the production of a reliable gold standard. To identify the cost of IE-inflicted information loss, a legal expert designed 5 multiple choice questions for four EULAs (e.g., 20 in total). All questions are related to `Permission`, `Prohibition` & `Duty`. In the last step 6 volunteers from the university campus (postgraduate students and staff) were required to answer these questions using two methods: i) reading the EULA in full text and ii) utilizing *EULAide*. The results are briefly presented in section 5.

4.2 *KnIGHT*

Privacy policies are legal documents stipulating how companies will gather, manage and process customer data. They are legally required for any service that uses, maintains or discloses data that can be used to identify an individual. In contrast to EULAs, privacy policies must comply with a smaller set of legislation, i.e., data protection laws. This focus enables us to perform more specific analysis and check compliance against specific data protection regulation. For such contractual agreements, we employ a deep learning approach utilizing an existing corpus. The OPP-115 dataset is divided into paragraphs, each of which includes annotations from three legal experts. There are two types of annotations: at the top level each paragraph is labeled with one (or more) pre-defined classes; and at a lower level a class may contain specific attributes. For example, the top level category “User Choice/Control” can be narrowed down to: choice type, choice



Permission

CREATE ACCOUNT

UPLOAD CONTENT

OFFERED

USE SERVICES , USES IN POLICY , USE SOFTWARE , USING SERVICES

1 You may use our Services only as permitted by law, including applicable export and re-export control laws and regulations

2 You can find more information about how Google uses and stores content in the privacy policy or additional terms for particular Services

3 gives you a personal, worldwide, royalty-free, non-assignable and non-exclusive license to use the software

Duty

FOLLOW POLICIES

LOOK

DISCONTINUE USE OF SERVICE

If you do not agree to the modified terms for a Service, you should discontinue your use of that Service

more info

Prohibition

LEASE PART OF SERVICES

USE CONTENT FROM SERVICES , USE BRANDING , USE SERVICES

1 You may not use content from our Services unless you obtain permission from its owner or are otherwise permitted by law

2 terms do not grant you the right to use any branding or logos used in our Services

3 Do not use such Services in a way that distracts you and prevents you from obeying traffic or safety laws

more info

Fig. 2: EULAide Platform Web interface showing the permission, duty and prohibition clusters for a user provided EULA (Google terms of service)

scope, personal information type, purpose. We are working on training *KnIGHT* with OPP-115 and extract top level classes and lower level attributes from a policy. Having this structured information from privacy policies, we are able to: 1) check their completeness according to the legislation; 2) measure their risk factor by analyzing the values of attributes; 3) map the top level categories to data protection laws for more advanced users (legal experts, data officers, etc.).

The first two goals target regular users as the intended audience, whereas the third one is more suitable for experts. Although the OPP-115 consists of policies defined by American companies, the top level categories can be mapped to GDPR. For instance, the category “First Party Collection/Use” is related to Article 13, `‘Information to be provided where personal data are collected’`, or “User Access, Edit & Deletion“ category can be linked to Article 16 & 17 (`‘Right to Rectification/Erasure’`). The mappings can be as general as a whole article or as detailed as a specific paragraph. For the evaluation of *KnIGHT*, the first two directions will be assessed using the OPP-115 as a gold standard, while the mapping accuracy should be assessed by experts.

5 Results

In this section, we will only discuss results from experiments seeking to evaluate the support provided by *EULAide* to make sense of legal agreements (terms of use). The evaluation of our privacy-policy (training data-based) approach is still in planning stage. In order to measure the performance of the OBIE pipeline, we have used our compiled gold standard based on the manually-annotated examples (excluding the 10% disagreement in the IAA exercise). The evaluation results are shown in tables 2a & 2b. The ontology enhancement was a feedback cycle during which we improved domain-specific coverage by adding additional instances (around 50), with the support of a legal expert. Considering the complexity of EULAs and the 90% agreement observed between human annotators, the results indicate that our OBIE method yields useful results and is feasible. The (90 - 72)% information loss comes from the incomplete set of grammar rules and ODRL instances coverage. Expanding the ontology with more concepts will allow us to define more rules and will eventually increase the system accuracy.

	Precision	**Recall**	**F1**
Permission	0.75	0.56	0.64
Prohibition	0.89	0.47	0.61
Duty	0.73	0.43	0.54
Overall	0.79	0.49	0.6

(a) without ontology enhancement

	Precision	**Recall**	**F1**
Permission	0.74	0.75	0.74
Prohibition	0.89	0.63	0.74
Duty	0.66	0.67	0.67
Overall	0.75	0.68	0.72

(b) with ontology enhancement

Table 2: Evaluation of OBIE against the goldstandard (%)

Tables 3a & 3b present results from the previously described *EULAide* usability experiment. Phase1 and Phase2 indicate different phases when answering the multiple-choice questions. In the first phase participants read EULAs (either in their full text or utilizing *EULAide*) and answered the questions using their memory. In the second phase they were allowed to use search tools for unanswered questions. The rationale behind this setting was to recreate the baseline method for users to check and read policies without any tool. Thus, we sought to identify how well regular people can remember policies and how fast they can search for information in an EULA. In practice, when one is agreeing with terms, this process should be followed so as to be aware of the rights and regulations. Our results verify our initial hypotheses, i.e., even though *EULAide* is effected by a (12 - 1.5 = 10.5)% information loss, it considerably saves time and effort spent by users to arrive to a similar level of understanding. Finally it should be stated that although due to funding restrictions the number of selected EULAs and participants was the bare minimum required for an experiment of this kind, the results were sufficient to indicate the value in extending and improving our approach.

	Reading	Phase1	Phase2
EULA Full Text	1185	75	152
EULAide	315	72	77

(a) Average time (In Sec.)

	Correct	Incorrect	Unanswered in Phase1		
			Phase2 Correct	Phase2 Incorrect	Phase2 Unanswered
EULA Full Text	67	8	18.5	5	1.5
EULAide	62	15	6.5	4.5	12

(b) Average percentage of questions results (%)

Table 3: Multiple-choice question answering results by 6 participants

6 Discussion

This thesis tackles the important issue of difficult-to-read legal documents and investigates automated methods for the benefit of end-users. The experiments conducted confirm the complexity of the task and the subjectivity of human judgment. Somewhat counter-intuitively, we observe that agreement between the experts is generally harder to achieve than between average users. This is probably due to the experts' higher understanding and ability for a more critical inspection of legal texts. A constant non-technical challenge in our efforts is to attain commitment from legal experts on a voluntary basis. Despite the challenges and difficulties, our results so far indicate that NLP techniques combined with OBIE and ML can be very useful to support legal text comprehension and that with sufficient funding broader experiments can be carried out.

Acknowledgments

I would like to thank Simon Scerri, Sören Auer and Jens Lehmann for their guidance and fruitful discussions during the development of this doctoral work.

References

1. Cabrio, E., Palmero Aprosio, A., Villata, S.: These are your rights. In: The Semantic Web: Trends and Challenges. pp. 255–269. Springer International Publishing (2014)
2. Costante, E., Sun, Y., Petković, M., den Hartog, J.: A machine learning solution to assess privacy policy completeness: (short paper). In: Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society. pp. 91–96. WPES '12, ACM, New York, NY, USA (2012). https://doi.org/10.1145/2381966.2381979
3. Cunningham, H., Maynard, D., Tablan, V.: JAPE: a Java Annotation Patterns Engine (Second Edition). Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield (November 2000), `http://www.dcs.shef.ac.uk/~diana/Papers/jape.ps`
4. Guntamukkala, N., Dara, R., Grewal, G.W.: A machine-learning based approach for measuring the completeness of online privacy policies. 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA) pp. 289–294 (2015)
5. Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K.G., Aberer, K.: Polisis: Automated analysis and presentation of privacy policies using deep learning. CoRR **abs/1802.02561** (2018)
6. Obar, J.A., Oeldorf-Hirsch, A.: The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. Information, Communication & Society pp. 1–20 (2018)
7. Wilson, S., Schaub, F., Dara, A.A., Liu, F., Cherivirala, S., Leon, P.G., Andersen, M.S., Zimmeck, S., Sathyendra, K.M., Russell, N.C., Norton, T.B., Hovy, E.H., Reidenberg, J.R., Sadeh, N.M.: The creation and analysis of a website privacy policy corpus. In: ACL (2016)
8. Wimalasuriya, D.C., Dou, D.: Ontology-based information extraction: An introduction and a survey of current approaches. J. Inf. Sci. **36**(3), 306–323 (Jun 2010). https://doi.org/10.1177/0165551509360123, `http://dx.doi.org/10.1177/0165551509360123`