CERIAS Tech Report 2007-102

Reconciling Privacy Policies and Regulations: Ontological Semantics Perspective

by Olga Krachina, Victor Raskin, Katrina Triezenberg
Center for Education and Research
Information Assurance and Security
Purdue University, West Lafayette, IN 47907-2086

Reconciling Privacy Policies and Regulations: Ontological Semantics Perspective

Olga Krachina[1], Victor Raskin[1], and Katrina Triezenberg[1]
[1] CERIAS, Purdue University
686 Oval Drive
W. Lafayette, IN 47907-2086, USA
{okrachin, vraskin, kattriez}@purdue.edu

Abstract. How well the privacy policy follows a regulation is one of the current concerns of the user. Such a task can be accomplished by directly querying the policy statement with the regulation text. Automation of the process requires an expressive meaning-based framework for Natural Language Processing (NLP). This paper discusses the Ontological Semantics approach to the issue of verifying compliance and illustrates the potential of utilizing the framework in the domain of Privacy management for NLP-related tasks. As an example a section from BCBS and corresponding HIPAA regulations are used.

Keywords: Ontological Semantics approach, ontology, inference, Text-Meaning Representation, Privacy Policy, Natural Language Processing.

1 Introduction

Although formal language policies are being developed and deployed, it is unlikely that they will replace natural language privacy policies (PP) in the foreseeable future [1]. While the deployment of the privacy practices described in a particular policy takes place on the level of a formal language, certain tasks can and should be handled on a higher level, i.e. via a natural language processing (NLP) system. The motivation for the following division of labor between formal and natural language arises for two reasons. First, an intermediate step of natural-language to formal-language translation is required to work with the privacy policy on the level of a formal language; in the process, loss of potentially useful information is unavoidable if the formal language is not able to support the entire content. The notion of a formal language discussed in this context implies a domain-specific construct, thus placing apparent limitations on the range of the input to be processed. In such a case, an extensive research is necessary to account for all or most of the possible inputs for a specific domain.

The second reason is a logical consequence of the issue discussed above, namely, formal languages by design are not able to achieve the same level of expressiveness as a large-scale NLP system. If such an NLP system is designed to implement domain-independent mechanisms to process natural language input, it should comprise several extensive knowledge resources built and updated overtime.

One of the central tasks in the field of privacy management is the issue of policy consistency with the outlined regulations. This paper will demonstrate how this issue

is handled from the perspective of one specific NLP system.

2 Ontological Semantics in the PP Domain: Motivation

There is an apparent need in a comprehensive NLP system. Such system should be able to handle complex natural language structures; this implies disambiguation of the input not only on the lexicographic level but also on the semantic level, integrating the deduction mechanisms as well as those of induction. In other words, an NLP system has to employ mechanisms based on rather expressive semantics. One of the previously considered frameworks was P3P. There have been attempts at extending formal semantics of P3P and the underlying formal language APPEL [7].

Such attempts were directed towards creating a language that "avoids the APPEL's pitfalls but preserves the desirable functionalities in APPEL" [7]. The main objective of APPEL is to allow users to import preference rulesets created by other parties and to transport their own ruleset files between multiple user agents [7]. Among major drawbacks of APPEL is its syntax-based design, i.e. a new ruleset is generated for two P3P policies with the same meaning but different syntactic structure.

The new desiderata for an enhanced APPEL-like language included expressive power and semantic consistency as outlined in [7]. On the one hand, it appears reasonable to try and create a language exclusively for the purpose of analysis of privacy policies; on the other hand, it may be more efficient to extend a currently existing NLP system to the domain of PP. The framework proposed here for the task of handling problems relating to NLP is Ontological Semantics (Nirenburg and Raskin 2004 [9]). Its efficiency has been demonstrated in such applications as automatic translation, question-and-answering, information retrieval, text summarization, Internet search, and other NLP systems, largely due to the fact that it is not an *ad hoc* solution for a specific NLP problem but rather a comprehensive and systematic approach. Thus, in order to handle the domain of the privacy policy management within Ontological Semantics (OntoSem), the extension of existing resources may be the only requirement.

The rest of the paper is structured as follows: Ontological Semantics will be introduced, then a specific example will demonstrate mechanisms employed in evaluating compliance of an entity with a corresponding regulation, and finally, evaluation metrics will conclude the discussion.

2.1 OntoSem Framework

Text analysis in OntoSem relies on the results of several preprocessing modules performing lexicographic, morphological and syntactic analysis. The results of such a pre-semantic analysis, along with the support of the knowledge resources, contribute to the word-sense disambiguation and the establishment of semantic dependencies. Such dependencies are expressed in a basic Text-Meaning-Representation (TMR), which are further extended with more sophisticated dependencies, such as preconditions, effects, and complex events, or scripts, and also with modalities, time, aspect, and other parameters, if necessary to arrive at an exhaustive meaning representation

of the input text. It is worth mentioning here that there is no one-to-one correspondence between a TMR and a sentence. It is often the case that a single TMR incorporates several syntactic units. TMR is the central construct of OntoSem as it underlies all text processing applications. It is an absolute knowledge representation module in a sense that TMR synonymy/paraphrase is not possible under OntoSem framework.

The OntoSem knowledge resources include the universal language-independent ontology (a large tangled hierarchy of concepts), language-specific lexicons, onomasticons (lexicons of proper names), and the Fact DataBase (FDB) containing instances of events occurred in the past processing of natural language texts, i.e. FDB is essentially a collection of TMRs. For further information on OntoSem see [8] and/or ontologicalsemantics.com.

2.2 More on the Goal of Compliance

As noted previously, the lack of a standard way to check compliance serves as one of the main motivating factors in finding a solution outside of the domain of formal languages for privacy policies, since their expressiveness is justifiably limited at the expense of other important factors such as the ease of use, etc [7]. Specifically, if the issue is approached from the perspective of natural language, its formulation can be reduced to the NLP question-and-answering problem. In other words, a PP notice can be queried with the regulation specifications.

One of the major hindrances in using an NLP system has been a necessity for "deep semantics" and inference capabilities and the ubiquitous "fear of semantics," due largely to the scarcity of adequately prepared semanticists. Those fears aside and the extended knowledge resources being available, this is still not an easily attainable goal, because such NLP system should be able to handle not only inductive and deductive reasoning but also reasoning under insufficient information or incomplete system resources. This research concerns itself with demonstrating how Ontological Semantics-based mechanisms handle inference problems.

3 Inference Processing with Ontological Semantics: Example

In Ontological Semantics, the TMR is the basic block for inference processing. Initially, the TMR corresponding to the query and the input text is constructed: TMRQ and TMRI, respectively. The Question-and-Answering mechanism is realized as a matching process of (see Section 3.1) between TMRQ and TMRI.

The following example is taken from HIPAA Section 164.522 on Rights to request privacy protection for protected health information and Section 164.528 on Accounting of disclosures[3]. This example not only illustrates (OntoSem) inference processing but also demonstrates the representation of a typical statement of the domain in Ontological Semantics.

Accounting of Disclosures:
164.522:

(iv) Individual Rights. The notice must contain a statement of the individual's rights with respect to protected health information and a brief description of how the individual may exercise these rights, as follows:

(E)[1] The right to receive an accounting of disclosures of protected health information as provided by §164.528

164.528:

An individual has a right to receive an accounting of disclosures of protected health information made by a covered entity in the six years prior to the date on which the accounting is requested, except for disclosures:

< …>

i. that occurred prior to the compliance date for the covered entity.

And the corresponding entry in BCBS PP [4] is as follows:

Right to Request an Accounting of Disclosures. You have a right to receive a list of certain instances in which we or our business associates disclosed your PHI for purposes other than our treatment, payment or health care operations and certain other activities. You are entitled to this accounting of disclosures for the six years prior to the date you make the request, but not for disclosures made before April 14, 2003.

Thus, relevant portions of TMRI are given in Listing (1), while Listing (2) is the selected portions of TMRQ. Indexing within FDB does not coincide with the indexing of the TMRI, since TMRI is not being committed to the FDB by default, i.e. the extension of FDB is controlled by the inference mechanism. Another thing to note is the proposition head REQUEST-INFORMATION which indicates the fact that the TMR is a representation of a query, i.e. TMRQ.

Listing 1. TMR of BCBS. Right to Request an Accounting of Disclosures.

```
TRANSFER-OBJECT-1
     BENEFICIARY                    HUMAN-1
     THEME                          LIST-1

LIST-1 (list of certain activities but not all)
     DESCRIBES                      INFORMATION-MANAGING-ACTIVITY-3
     THEME-OF                       TRANSFER-OBJECT-1
     SET                            TYPE(INFORMATION-MANAGING-ACTIVITY)
                           ELEMENTS(INFORMATION-MANAGING-ACTIVIITY-3)
                                    COMPLETE(NEG)

INFORMATION-MANAGING-ACTIVITY-3
     AGENT                          ORGANIZATION-2
                                    THIRD-PARTY
     HAS-EVENT-AS-PART              INFORM-1
     PURPOSE                        SET(PAY-1
                                    TREAT-ILLNESS
                                    SERVICE-EVENT) COMPLETE(NEG)
INFORM-1
     AGENT                          ORGANIZATION-1
     BENEFICIARY                    THIRD-PARTY
     THEME                          PHI
     PUBLIC-ATTRIBUTE
            time_1.value            PRIVATE
```

[1] Skip sections A-D

```
            time_1              < speech-act.time
            time_2.value        PUBLIC
            time_2              > = speech-act.time
      TIME                      |<16><04><2003>|
                                < REQUEST-INFORMATION-1.TIME
            range               ::years:: <= 6
            frequency           default

LEGAL-RIGHT
      THEME                     TRANSFER-OBJECT-1
      AGENT                     HUMAN-1
      BENEFICIARY               HUMAN-1
```

A significant part of the process of verifying compliance lies in the construction of the query itself. There are two basic types of queries: 'wh'-type and 'yes-no'. The latter are more suitable for the task of determining compliance. This is due to the style of regulation documents: these texts as a rule indicate the obligative nature of existence of certain provisions or statements; such a structure, in turn, allows instantiation of REQUEST-INFORMATION field with THEME of main event in the TMRQ or relation as it turned out to be in this example (Listing 2).

Additionally, TMRQ contains a *LEVEL* slot, which indicates the desired depth of the resulting answer and takes on integer values. In other words, the mechanism accounts for cases where an exhaustive answer is not necessary. The motivation for this parameter stems from two reasons. On the one hand, the answer to a wh-query is a chain of dependencies (error estimation with a positive-valued response for yes-no queries) starting from the target variable itself and ending with the last relevant head of the frame, which in some cases is not necessary and is a waste of time and resources; on the other hand, introducing this parameter is an efficient step towards the goal of automation of query construction, i.e., instead of constructing a new query TMR from already existing TMRI, only an instantiation of REQUEST-INFOMRATION (and MODALITY in certain cases) fields with appropriate event/object focus are sufficient to formulate a query which is further constrained through the LEVEL parameter. In this example, the LEVEL parameter is set to its maximum. Furthermore, appropriate TMRQs can be constructed based on the filler (in case of a relation, a corresponding event or an object associated with it) of the slot 'scope' of the deontic modality [9], i.e., the modality that expresses obligation.

Listing 2.

```
MODALITY-1
      type                DEONTIC
      value               1.0
      time                > speech-act.time
      scope               DOCUMENT-1.TEXTUAL-RELATION
      attributed-to       ORGANIZATION-2

DOCUMENT-1
      DESCRIBES           LEGAL-RIGHT-1
                          RULE-OF-CONDUCT
      AUTHORED-BY         ORGANIZATION-1
      LOCATION            WEB
      TEXTUAL-RELATION    SENTENCE
```

```
SENTENCE
     DESCRIBES             LEGAL-RIGHT-1
     TEXTUAL-RELATION      DOCUMENT-1

AUTHORIZE
     AGENT                 ORGANIZATION-2
     THEME                 LEGAL-RIGHT-1
     BENEFICIARY           HUMAN-1

LEGAL-RIGHT-1
     AGENT                 HUMAN-1
     BENEFICIARY           HUMAN-1
     THEME                 INFORMATION-MANAGING-ACTIVITY
     CARDINALITY           PLURAL
     SET                   MEMBER-TYPE(LEGAL-RIGHT)
                           ELEMENTS ((LEGAL-RIGHT-2)) COMPLETE (NEG)
LEGAL-RIGHT-2
     THEME                 TRANSFER-OBJECT

TRANSFER-OBJECT
     THEME                 LIST-1
     AGENT                 ORGANIZATION-1
     BENEFICIARY           HUMAN-1

LIST-1
     DESCRIBES             INFORM-1

INFORM-2
     AGENT                 ORGANIZATION-1
     THEME                 INFORMATION-1
     PUBLIC-ATTRIBUTE
            time_1.value   PRIVATE
            time_1         < speech-act.time
            time_2.value   PUBLIC
            time_2         > = speech-act.time
     BENEFICIARY           THIRD-PARTY
     TIME                  < OBEY.time
         range             ::years:: <= 6
         frequency         default

OBEY
     AGENT                 HUMAN-1
     THEME                 RULE-OF-CONDUCT

RULE-OF-CONDUCT
     DESCRIBES             INFORMATION-MANAGING-ACTIVITY-1

INFORMATION-MANAGING-ACTIVITY-1
     AGENT                 HUMAN-1
     THEME                 PHI

INFORM-1
     AGENT                 ORGANIZATION-1 (BCBS)
     THEME                 INFORMATION-1 (PHI)
     SET                   MEMBER-TYPE(INFORM)
     ELEMENTS              (INFORM-2) COMPLETE (NEG)
     PUBLIC-ATTRIBUTE
            time_1.value   PRIVATE
            time_1         < speech-act.time
            time_2.value   PUBLIC
```

```
            time_2          > = speech-act.time
      BENEFICIARY           THIRD-PARTY

  ORGANIZATION-1
      HAS-NAME              BCBS
      AGENT-OF              INFORM-1
      AUTHOR-OF             DOCUMENT-1

  ORGANIZATION-2
      HAS-NAME              HIPAA
      AGENT-OF              AUTHORIZE
      AUTHORITY-ATTRIBUTE   1.0

  REQUEST-INFORMATION
      THEME                 LEGAL-RIGHT-1
      LEVEL                 <ALL>
```

The query described in Listing 2 corresponds to the natural language statement: "Does the PP contain a statement of individual rights *as described in HIPAA regulation*". Typically, a 'yes-no' query involves an event, however, it is not limited to just events as long as it is formulated in a way compatible with the selectional restrictions on the filler of the THEME slot, i.e., the RANGE values of ontology-slot.

3.1 Inference Process: Brief Overview

The algorithm governing inference process in Ontological semantics is given in Listing 3 below: it is comprised of four basic procedures: FDB search, direct TMR matching, TMR expansion, and construction of PREMISE-SET. FDB search is invoked in case of wh-queries if a proper-name (object or event) is present. It is the least computationally expensive portion of the algorithm; a case statement is used to express the three possible outcomes of the search and each of them is handled appropriately--for a more detailed example see [5]. Direct TMR matching is a more general procedure triggered in case a proper-name is missing from FDB or in case of 'yes-no' queries where FDB search is not invoked, but rather matching of proposition in INFORMATION-REQUEST THEME of TMRQ and a proposition head in TMRI is performed. It is denoted as 'direct' because propositions are matched without utilizing any of the ontological resources. TMR expansion and PREMISE-SET are triggered in case of 'yes-no' queries, i.e., when the control is transferred to the "else" statement.

Listing 3

```
TARGET = THEME-OF REQUEST-INFORMATION
IF(((TARGET := OBJECT) || (TARGET := EVENT))
       && (HAS_NAME (non-empty)))
  INFO = SEARCH_FDB(TARGET)
   case_1: single entry
           RETURN INFO
   case_2: multiple entries
       do REFINE_VARIABLE(TARGET)[2]
    case_3: no entry
       do DIRECT TMR MATCHING
```

[2] A discussion of FDB search-related specifics is omitted--for details see [5].

```
       APPEND_TO_FDB
       RESULT = TARGET
ELSE
     DIRECT_TMR_MATCHING (TARGET)
     DO TMR_EXPAND(PREMISE_SET)
     RESULT = EVALUATE(PREMISE_SET)
END IF
do ERROR_ESTIMATION
RESULT = RESULT + ERROR
RETURN (RESULT)
```

3.2 Inference Processing: Example

As the first step, the direct TMR matching procedure is invoked. It is assumed that, in the domain of PP, this particular procedure would be frequently used, especially in cases where PP aligns or is expected to align closely with the regulation due to the genre in which both texts are written. Once the THEME field of INFORMATION-REQUEST in TMRQ is matched to a proposition in TMRI, all its slots are being matched against those of the concept in TMRQ to the degree indicated in the LEVEL field. In other words, two dependency chains starting with THEME of REQUEST-INFORMATION are compared against each other and error estimates are attached in case of discrepancies or mismatches.

For the given example, mismatch occurs at a point of the LIST-1 DESCRIBE case role filler: INFORM-1 and INFORMATION-MANAGING-ACTIVITY-3 in TMRQ and TMRI, respectively. These mismatched items as well as the established dependencies are fed to the PREMISE-SET.

PREMISE-SET is a procedure that explores the OntoSem hierarchical structure; it updates two groups of relations in TMRQ: the case-roles and subsumption relations (see [5]). The system will attempt to establish a Most Common Intermediate Node (MCIN) by looking for a proposition in TMRI which results in the closest MCIN with the mismatched entry from the TMRQ. In this process, INFORM-1 will be found before MCIN is established, since INFORM appears as a proposition head and qualifies as a potential candidate for establishing MCIN. Its relation to the mismatched entry INFORMATION-MANAGING-ACTIVITY is established in the backtracking fashion.

At this point, PREMISE-SET contains the relation of the mismatched entries; further inconsistencies or discrepancies of the two TMRs are evaluated and passed into the result-formulating component in corresponding format.

The answer-formulating component collects error estimates as well as a reference dependency chain with respect to which the errors were estimated. The final answer is expressed as a natural language corresponding statement of the reference dependency chain with associated errors.

3.3 Error Estimation

Given the complexity of OntoSem text analysis, development of evaluation metrics is not an easy task: only preliminary efforts in this direction are outlined below.

The overall error assignment is based on the total number of tokens in the original query and the number of correctly inferred tokens [5]. Error estimation for each token

proceeds differently depending on the type of mismatch; for cases of discrepancy in proposition heads, the *underspecification* error is calculated as the distance from the MCIN to the node in TMRI and normalized by the distance from MCIN to the corresponding target node in TMRQ. In case of *overspecification* in the TMRI node, i.e., a case where the target node in TMRQ is less restricted than the corresponding node in TMRI, error estimation is not considered.

The same exact technique cannot be applied in case of mismatched properties as it is not particularly informative, e.g. PURPOSE and EXPERINCER are both children of CASE-ROLE, however, semantically, their meaning is very different. In fact, there may not be meaningful error estimation when property mismatches occur. In such cases a relational or attribute error, E(r) or E(a), respectively, with the specific instance, is passed on to the answer-formulation component. The Relational or attribute error are not generally gradable, and the severity of such errors is determined by the user agent.

In the discussed example, the main discrepancy adding to the significance of alignment verification occurs in TIME field; in particular, there is an extra value associated with INFORM.TIME. It will be reflected in the result component. Additional information specified in TMRI that is not present in TMRQ has a restricting effect on the answer to the query and cannot be accounted for in a systematic way, however, it is reported as a part of the answer as E(+).

The error estimation for the example discussed in this paper is calculated below:

- total number of tokens to match: 9
- number of tokens successfully matched: 7

$\therefore$ probability of error = $1 - 7\times1/9 + 1/9\times E(r) + 1/9\times E(+)$, where E(r) is a relational error and E(+) is error due to extra information, which are set to constant value zero, hence the total error with respect to the reference dependency chain in Listing 4 is 2/9.

Listing 4

```
LEGAL-RIGHT-1 (SET (LEGAL-RIGHT-2))
LEGAL-RIGHT-2  (THEME  (TRANSFER-OBJECT))
TRANSFER-OBJECT (THEME (LIST))
LIST (DESCRIBES (INFORM-1))
INFORM-1 (SET (INFORM-2))
INFORM-2 (TIME (::YEARS = 6::))
INFORM-2 (THEME (PHI))
PHI (DESCRIBES (HUMAN))
THIRD-PARTY(BENEFICIARY-OF (INFORM-2, INFORM-1))
ORGANIZATION-1 (AGENT-OF (INFORM-1, INFORM-2))
E(+):
  INFORM-2.TIME                    |<16><04><2003>|
                                   < REQUEST-INFORMATION-1.TIME
              range                ::years:: <= 6
              frequency            default
E(r):
      LIST (DESCRIBES (INFORMATION-MANAGING-ACTIVITY))
      INFORMATION-MANAGING-ACTIVITY (HAS-EVENT-AS-PART (INFORM-1))
```

4 Conclusion and Future Research

Many problems that arise in the domain of PP management are due to the inherent ambiguity of natural language. Thus, it is expedient to look for solutions to such problems from the perspective of systems that incorporate adequate resources.

This paper focused on using the OntoSem framework in the domain of Privacy Policy management; specifically, it investigated the problem of establishing or checking PP compliance with regulations. The advantage of using OntoSem as an underlying framework is in the structure and interaction of its resources as well as in the fundamentals of knowledge representation. Given that the problem of policy compliance can be reduced to a question-and-answering problem, solution lies in the development and enhancement of NLP inference methods.

In order to accommodate the domain under OntoSem, current knowledge resources were extended: new lexical entries (500) as well as ontological concepts (50) were acquired (see [6]). FDB will be populated with the TMRs from domain with the processing of new texts. Even though, at the core of the inference processing, there lies a high-quality TMR, there still exists a degree of ambiguity, which is expressed in terms of error probabilities.

Future research will focus on exploring the application of OntoSem inference capabilities in the domain of Privacy Policies management--specifically, abstracting out a standard set of queries that are necessary to determine compliance based on a particular category of regulation document. Such a task would necessarily involve a development of a ranking system of the queries and will result in gradable outcomes. Implementation of this idea may require a development of a new framework particular query type. OntoSem-specific future research goals include the improvement and evaluation of automated TMR construction [cf. 8] and extension of current knowledge resources and further improvement and increasing automation of acquisition methodologies [10].

References

1. Anton, A., Baumer, D., Bertino, E., Dark, M., Li, N., Proctor, R., Rappa, M., Raskin, V., Vu, K., Yu, T. 2004. CyberTrust Proposal, National Science Foundation.
2. Blue Cross Blue Shield Privacy Policy of North Carolina.
3. Health Insurance Portability and Accountability Act. 1996.
4. KBAE. Knowledge Base Acquisition Editor.
5. Krachina, O. 2006. Ontology-Based Inference Methods. Proceedings of MCLC 2006, Urbana-Champaign, IL: University of Illinois, May 2006.
6. Krachina, O. 2005. Role of Ontological Semantics in Handling Privacy Policies. Poster. 2005 CERIAS Annual Research Symposium, West Lafayette, IN. CERIAS TR-XX-2005.
7. Li, N., Yu, T., Anton, A. A Semantics-Based Approach to Privacy Languages.To appear in *Computer Science and Systems Engineering Journal*.
8. Nirenburg, S., Beale, S., McShane M. 2004. Evaluating the Performance of the OntoSem Semantic Analyzer.
9. Nirenburg, S., Raskin, V. 2004. Ontological Semantics. Cambridge, MA: MIT Press.
10. Triezenberg, K. E. 2006. The Ontology of Emotion. An Unpublished Ph.D. Thesis, Linguistics, Purdue University.