Bridging the Semantic Gap between Sensor Data and High Level Knowledge

Marjan Alirezaie and Amy Loutfi
Cognitive Robotic Systems Lab, Center for Applied Autonomous Sensor Systems (AASS)
Dept. of Science and Technology, Örebro University, SE-701 82, Örebro, Sweden

1 STAGE OF THE RESEARCH

Ubiquitous devices such as sensors in weather stations, health-care centres and generally wired/wireless sensor networks enable massive amounts of observations called big data increasing with the rate of 2.5 quintillion bytes per day (Hilbert and López, 2011). This amount of online data provides a large potential to have a deeper and better understanding of the world around us. These data which are mostly time-series signals must be interpreted and represented in a manner that is compatible for humans. Typically, this interpretation is automated by using complex data driven analysis methods. The output of such methods can still contain inaccuracies due to the fact that low level sensor data is subject to shortcomings due to selectivity, uncertainties and errors. In other words, while data analysis techniques can provide interpretation of data in different forms of event detection, more is required for inducing their meanings to the users in the context of multiple sensor monitoring.

Furthermore, the Linked Data Cloud including axioms in different domains such as geographic, government, media and life sciences is growing up so much that it reached to 31 billion RDF triples in 2011 from over two billion RDF triples in 2007 (Heath and Bizer, 2011). This interleaved deluge of axioms that is increasingly becoming connected into a rich network of other data sources can be regarded as a source of knowledge for tasks demanding the common-sense knowledge. In addition, the machine processable language of this knowledge cloud makes it suitable for an automatic way of reading and on the whole using its contents.

Infusing high level knowledge into the observations of the environment generated by sensors helps to enhance data interpretation and make ”sense of sensor data”. However, data driven processes manipulating sensor readings are not able to automatically consider the wealth of high level knowledge for the integration. In other words, what is needed is a method which is able to automatically bridge the *semantic gap* between qualitative high level knowledge and quantitative low level data which are inherently difficult to interpret. This work is part of an ongoing effort to specifically address a particular instantiation of the semantic gap special and unintuitive data coming from particular chemical sensors measuring gases or odours and the knowledge that humans have about odours (Loutfi et al., 2001; Loutfi, 2006). The work developed in this thesis will also examine general methods that can be applied to various domains involving continous time series data. This paper describes the approach used to work towards this goal. The paper outlines the main contributions of the work, details the progress so far after two year of the thesis work, and provides an outline of planned activities.

2 OUTLINE OF OBJECTIVES

The objective of this research is to employ abductive (non-monotonic) reasoning to automatically determine correspondences between sensor data (e.g. time series signals) and concepts in massive knowledge sources (e.g. Linked Open Data). Abductive reasoning will provide the ”best explanations” for observed behaviour in the sensor data and according to the principals of non-monotonic reasoning (Eiter et al., 1997), the ”best explanation” in this case is defined based on two parameters:

- Covering (covers all observed things)
- Minimality (reduces redundancy)

Given a set of rules, deductive reasoning whose inferring process goes from *effects* to *causes* has more observation-dependent answers (Pagnucco, 1996) whereas the abductive reasoning (*causes* to *effects*) considers all the rules having at least one observed premise's item. It means that abductive reasoning can hypothesize knowledge that is needed for the inferring process but is not necessarily available. This feature of abductive reasoning qualifies it for inferring

Alirezaie M. and Loutfi A. (2013).
Bridging the Semantic Gap between Sensor Data and High Level Knowledge.
In *Doctoral Consortium*, pages 21-27

from incomplete knowledge. More specifically, using this reasoning technique, the best explanations which are dependent to the current amount of facts can be *inferred* where the available knowledge is not complete.

Abductive reasoning as such entails controversial issues for which the following prerequisites are required:

- A formalisation which unifies the heterogeneous knowledge from knowledge sources (knowledge representation)
- A method to relax time complexity given an abductive reasoning framework satisfying covering and minimality conditions (efficiency of reasoning)

These prerequisites form the basis of the research problem. However, before going further into their technical details, it is worth stating the importance of formalisation of knowledge. Knowledge formalizing is about analysing a body of knowledge in order to translate it into a predefined language having its own vocabularies, notations and syntactical rules. In this way, the accurate studying of properties of concepts especially where the knowledge is expressed in qualitative terms than quantitative terms is improved (Balduccini and Girotto, 2010), and consequently the possibility of having better knowledge inference increases.

From the knowledge management point of view, defining a formalization and subsequently implementing an interpreter especially when the input knowledge is not homogeneous is a challenge. Addressing this challenge as well as the efficiency of the reasoning process are considered as two of the specific contributions of this research work.

3 RESEARCH PROBLEM

One of the research problems addressed in this thesis is the problem of encoding a formalisation for heterogeneous knowledge modelled in RDF/OWL. A formalization is generally defined based on a set of syntactical rules along with a set of notations. During a formalization process, the input which in this research is a set of concepts coming from heterogeneous knowledge sources is translated into a new language. On account of the ontological structure of the knowledge repositories, encoding an "RDF/OWL friendly" formalisation can speedup this process. For example, the concepts such as "object property" and "data property" which are directly recognisable from RDF triples can be immediately encoded into predicates (in a logic program) which show the specific relations among entities. In this way, no matter how the knowledge concepts are represented, the reasoner receives the set of statements, namely a logic program implemented within the notations of this formalisation. This encoding process indicates the necessity of an interpreter which has the task of converting a knowledge body formed in RDF/OWL triples into a logic program defined based on the standard of the formalization.

However, since in OWL/RDF conventions there is no limitation for labelling the classes, individuals and properties the formalization process can meet a concept modelled in distinct knowledge bodies but with different names. If these redundancies are not addressed, the formalization process instead of creating several logic rules which are atleast common in one predicate (or generally an atom), may build some independent rules.

In addition, because of computational factors such as incompatible hypothesis selection and satisfying the maximum plausibility, in general, abductive reasoning is an NP-hard problem (Bylander, 1991) and some heuristics are required to reduce its computational complexity. Since non-monotonic reasoning is inherently sensitive to adding new facts, as a solution, we can recognise the missing facts increasing the time complexity and create a searching term to be submitted through the knowledge sources. The second problem in this research is thus implementing a sub-process for the reasoner which by recognising the missing facts provides a guided search in order to reduce the complexity.

4 STATE OF THE ART

The common root of all works in integration of data in different levels of abstraction is in data fusion. The focus of fusion methods such as (Joshi and Sanderson, 1999) is on raw data consolidation for the sake of better data interpretation, however, without any integration with higher level of data (symbolic knowledge).

In order to consider fusing symbolic knowledge to numeric data, works such as (Loutfi et al., 2005; Coradeschi et al., 2013) have gained attention in AI fields related to robotics and physically embedded systems. The symbol grounding problem (Harnad, 1990) in general and the anchoring problem (Loutfi et al., 2005) in particular concentrate on the process of creating and maintaining the relation between a symbol chosen to label an object in the world and those data coming from sensors observing the same phys-

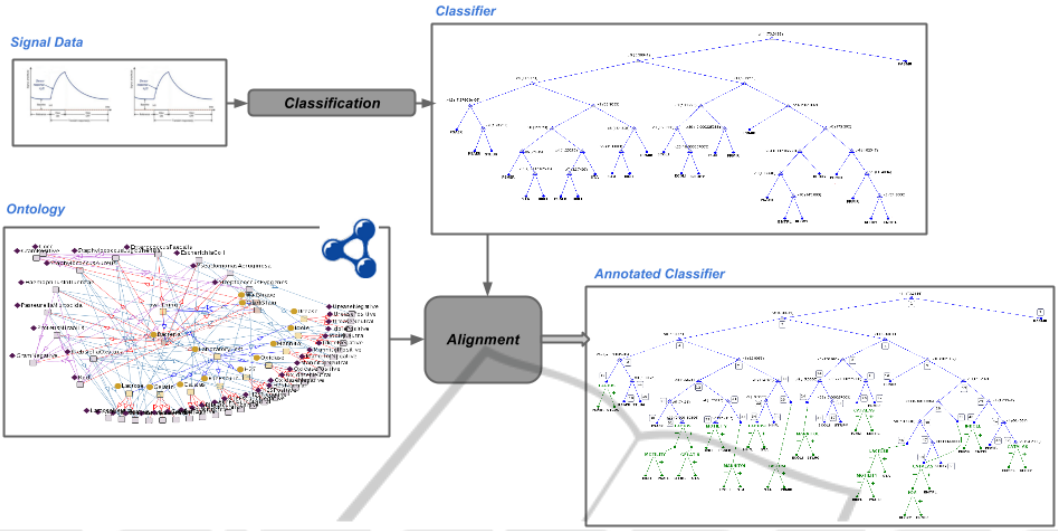

Figure 1: Ontology-Classification Alignment.

ical object in the environment. In most of works, for example (Melchert et al., 2007), the challenge is how to perform the anchoring in an artificial system and how to find relevant concepts related to symbols to improve the recognition process. In these efforts the association is done in two ways: grounding well-defined concepts in data (top-down process) and conceptualizing data that exemplifies the concept (bottom-up process). In these kinds of mapping we need to have the information about the objects mostly in forms of production rules which are manually (not automatically) modelled and are suitable for deductive reasoning in the environment.

The necessity of a posteriori model implied by the automatic knowledge acquisition approach has recently emerged in the area of sensor data processing. The work (Thirunarayan et al., 2009) applies abductive reasoning over sensor data which are interpreted based on predefined knowledge. Other works such as (Henson et al., 2011) and (Henson et al., 2012), model a system that makes it possible to infer explanations from an incomplete set of observations which are not necessarily sensor data. The reasoning framework in these works is based on Parsimonious Covering Theory (PCT) (Reggia and Peng, 1986). Nonetheless, since PCT in these works is translated into OWL, it is not able to provide an explanation containing more than one cause for the observations. Consequently, following the non-monotonic reasoning approach as the connector of two levels of represented data, we aim to encode a framework for automatizing this integration process as independent to the domain as possible.

5 METHODOLOGY

The process bridging the so-called semantic gap by aligning semantic knowledge to sensor data has proceeded in a three stage process. The final stage has emerged from discovery of the shortcomings in the previous two stages. Recalling the specific goal of this research work which is about odour sensor data, in this section, we describe the thesis work so far and the approaches used to address the overall aims of the thesis.

5.1 Improved Classification of Multivariate Data using Ontology Alignment

In this approach we consider a scenario where sensor data that is classifiable into well known categories from a data driven method is aligned to concepts which are part of a larger ontological structure. More specifically, this type of alignment which is recommended for situations where sensor data are unintuitive (e.g. electronic nose data) takes advantage of the classifiers' topology. Utilizing ontological alignment methods such as string and structural matching techniques, it finds relevant informative concepts from ontologies in order to resolve the misclassification. In this approach the ontological alignment techniques are used to align an ontology to a hierarchical classifier (i.e. decision trees) representing the features used in classification of labelled sensor signals. The output of the system is in effect a recommender system, and the reason for this is largely due to the type

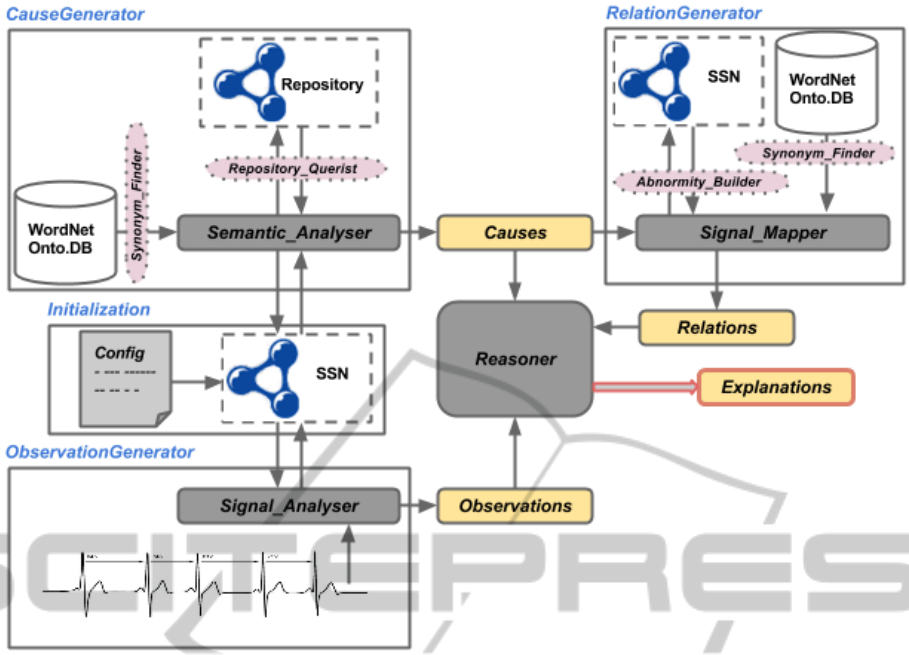

Figure 2: Data Streams Annotation/PCT Reasoner.

of available knowledge of the studied domain. The steps of this methodology whose outputs are shortly listed below:

- Classifying pre-processed sensor data (hierarchical classifiers)
- Localizing misclassified cases in the output of the classifier
- Aligning the classifier and the ontology to find similar parts
- Replacing candidate parts of the ontology with their counterparts in the classifier

In this solution, the focus of the alignment method is more on structure (inexact graph) matching than on semantic parts.

The specific data set used in the instantiation of this approach consists of time-series data from an electronic nose which is equipped with 22 sensors and "sniffs" the headspace of clinical blood samples containing 10 types of bacteria species. The objective is providing an estimate of the type of bacteria present in the sample. Extracting two descriptors from each signal, we eventually produce 44 feature values for the data set of 600 samples accompanied by a label list containing bacteria species names.

The result from the alignment is an improved classifier where recommendations are given to a user (expert) based on the interpretation of the sensor data that is done automatically. Figure 1 demonstrates the sequence of works in this approach.

More details on this work can be found in (Alirezaie and Loutfi, 2012). This work because of the shortcoming which was due to the lack of knowledge regarding the domain follows the structural (topological) techniques for the alignment process rather than the semantic analysis. Therefore, the concentration is mostly on string matching process between the labels of the concepts in the ontology and labels of the nodes in the classifier. However, for richer data sets measuring different properties of the environment where more knowledge are available, the alignment process can take advantage of the reasoning that provides the semantic analysis.

5.2 Reasoning about Sensor Data Annotations using PCT

An inherent part of solutions for the semantic gap problem is the ability to annotate the signals and in particular to annotate interesting events with plausible explanations. It is worth empowering the semantic gap filling process in terms of semantic analysis if a more meaningful multivariate data set is the target of the annotation task. The alignment method explained in Section 5.1 was suffering from the poor data in the sense that only one property of the phenomenon (odour) lacking high level knowledge was measured.

Developing the scenario towards having multiple sources of sensor data, we study the process of annotation. For instance, in a sensor network where a sensor is accompanied by others that are at the same

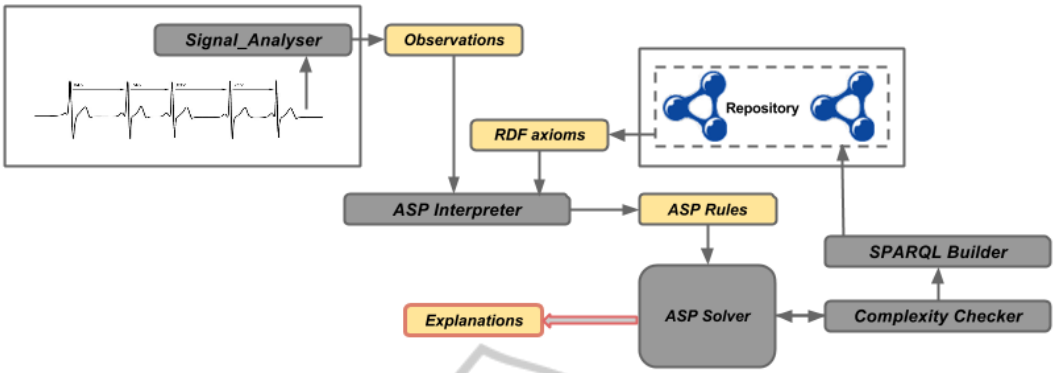

Figure 3: ASP Alignment technique.

time measuring different features of the environment, we can *reason* about their influence on each other.

Considering the aforementioned features of the abductive reasoning, we found it as a complementary technique for the knowledge retrieving task where the process needs to sift the best (most relevant) annotations. The alignment method of this approach is a type of abductive reasoner working based on Parsimonious Covering Theory (PCT) (Hilbert and López, 2011). Shown in Fig 2, the reasoner receiving three inputs, namely Observations, Causes and Relations finally results in Explanations which is considered as a set of annotations for what are observed over signals. The basis of this theory is on the Set theory in mathematics so that the reasoner calculates the power set of the Causes list to find the best explanation. In other words, the items of the final Explanations are subsets of the Causes list chosen based on the reasoner principles. According to this theory, the best explanation is defined within two criteria: Covering and Minimality. With the former criterion, the reasoner nominates those subsets of the Causes list whose items are related to all the observations detected in a particular segment of signals. Furthermore, the minimality criterion which is also called irredundancy considers the size of the aforementioned selected subset. In this way, the reasoner is able to choose those covering subsets of the Causes list that are minimal in terms of the cardinality.

To evaluate this work, we use multivariate data coming from medical sensors observing a patient suffering from several diseases as the ground truth against which the eventual explanations (annotations) of the reasoner are compared. This data set is a benchmark data set provided for 1994 AI in Medicine symposium submissions (Bache and Lichman, 2013). It contains 12-hours time-series ICU data (coming from 5 sensors) of a patient suffering from a set of disease. This package because of the richness of available online knowledge in medicine is well-suited for the evaluation task of this alignment technique where the semantics of concepts are involved in the reasoning process. However, we still have a long way to promote the alignment process. Although the PCT abductive model is to some extent analysing the semantics, its results are not yet declarative enough and are just copied from the labels of the concepts in ontologies. Further, we want to examine our original data set containing electronic nose data along with other types of data.

5.3 Reasoning about Sensor Data Annotations using ASP

Keeping the abductive reasoning approach as the alignment method, this framework hires an ASP based reasoner. Recalling the formalization phase, we want to model a framework that passes the body of knowledge into an interpreter which encodes the knowledge into a program familiar with the reasoner. As mentioned before, due to the expressivity of high level knowledge modelled by the domains' experts, we are aiming to exploit them to have more declarative interpretation for our observations. On the other hand, negation is a natural linguistic concept and extensively required when natural problems have to be modelled declaratively. Equipped with two different negation operators, weak (*not*) and strong ($\neg$) negations, as well as the disjunction operator ($\vee$) in the head of rules, the ASP based language, namely AnsProlog is known as a most suitable declarative language for knowledge representation, reasoning and declarative problem solving. More precisely, combining the two negation operators, answer set semantics provides the possibility for the reasoner to infer the natural language based and declarative explanations from the incomplete knowledge.

Therefore, in order to take advantage of the negation semantic, it is required to be considered in formalization and subsequently in reasoning phase. For

example, considering the meaningful operators of the answer set semantics, the interpreter is tasked to build an AnsProlog program from the RDF/OWL axioms and provides it for the reasoner which is aware of these notations, called ASP Solvers. Depicted in Fig 3, the knowledge sources which are modelled with RDF triples need to be aligned with the observations stating detected events in the environment. Since the ASP solver accepts AnsProlog clauses, the existence of an ASP Interpreter is indispensable. Given a set of RDF axioms, this interpreter creates ASP based rules. At this moment, the created rules are ready to be passed through the ASP Solver for inferring the best explanations.

However, this solver due to the amount of rules generated by the aforementioned component might be not efficient enough. For example, the inference process might be time consuming or even undecidable. In order to overcome these problems, the second main component is required. Being able to recognise the axioms absence of which raises the computational complexity, the Complexity Checker builds a SPARQL query based on the lack of knowledge and queries the repositories. Loosely speaking, this component close the loop of ontology-ASP Solver for the sake of relaxing the complexity by looking for highly required axioms over knowledge sources.

This approach will be evaluated with data coming from a small smart-kitchen equipped with a ZigBee networks including ZigBee sensors and an electronic nose that observe the environment. The data gathering phase is under process and the objective is annotation of the electronic nose data with the best explanation inferred by the reasoner.

6 EXPECTED OUTCOME

All three approaches are common in the final goal, namely annotating sensor data which can be counted as a solution for the semantic gap problem specifically for our electronic nose data. There are two parameters discerning among these models, the effectiveness in the sense of the time complexity and the expressiveness of final explanations. We will examine how multivariate data coming from sensors which are in company with electronic noses can promote the reasoning process in terms of creating the best explanations.

ACKNOWLEDGEMENTS

This work has been funded by the Swedish Research Counciln (Vetenskapsradet) under the project title *c*ognitive electronic noses.

REFERENCES

Alirezaie, M. and Loutfi, A. (2012). Ontology alignment for classification of low level sensor data. In Filipe, J. and Dietz, J. L. G., editors, *KEOD*, pages 89–97. SciTePress.

Bache, K. and Lichman, M. (2013). Uci-machine learning repository. Irvine, CA: University of California, School of Information and Computer Science.

Balduccini, M. and Girotto, S. (2010). Formalization of psychological knowledge in answer set programming and its application. *Theory Pract. Log. Program.*, 10(4-6):725–740.

Bylander, T., A. D. T. M. J. J. (1991). The computational complexity of abduction.

Coradeschi, S., Loutfi, A., and Wrede, B. (2013). A short review of symbol grounding in robotic and intelligent systems. volume 27, pages 129–136. Springer-Verlag.

Eiter, T., Gottlob, G., and Leone, N. (1997). Semantics and complexity of abduction from default theories. *Artificial Intelligence*, 90(12):177 – 223.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42:335–346.

Heath, T. and Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 1st edition.

Henson, C. A., Sheth, A. P., and Thirunarayan, K. (2012). Semantic perception: Converting sensory observations to abstractions. *IEEE Internet Computing*, 16(2):26–34.

Henson, C. A., Thirunarayan, K., Sheth, A. P., and , P. H. (2011). Representation of parsimonious covering theory in owl-dl. In *OWLED*.

Hilbert, M. and López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, 332(6025):60–65.

Joshi, R. and Sanderson, A. C. (1999). *Multisensor fusion : a minimal representation framework*. Series in Intelligent Control and Intelligent Automation. World Scientific, Singapore, London, Hong Kong.

Loutfi, A. (2006). *Odour Recognition using Electronic Noses in Robotic and Intelligent Systems*. PhD thesis, Örebro University, Örebro, Sweden.

Loutfi, A., Coradeschi, S., Duckett, T., and Wide, P. (2001). Odor source identification by grounding linguistic descriptions in an artificial nose. *Proc. SPIE Conference on Sensor Fusion: Architectures, Algorithms and Applications V*, 4385:273–282.

Loutfi, A., Coradeschi, S., and Saffiotti, A. (2005). Maintaining coherent perceptual information using anchoring. pages 1477–1482.

Melchert, J., Coradeschi, S., and Loutfi, A. (2007). Knowledge representation and reasoning for perceptual anchoring. *Tools with Artificial Intelligence*, 1:129–136.

Pagnucco, M. (1996). *The Role of Abductive Reasoning within the Process of Belief Revision*. PhD thesis, Basser Department of Computer Science, University of Sydney.

Reggia, J. A. and Peng, Y. (1986). Modeling diagnostic reasoning: A summary of parsimonious covering theory. *Comput Methods Programs Biomed*, 25(2):125–34.

Thirunarayan, K., Henson, C. A., and Sheth, A. P. (2009). Situation awareness via abductive reasoning from semantic sensor data: A preliminary report. In *CTS*, pages 111–118.