**Description du sujet:**

**Context**

The rise of deep learning has been made possible by the availability of large computing powers at affordable prices. Its performance has made it possible to solve problems that previously seemed out of reach, especially in the field of perception, such as computer vision and the analysis of natural language texts. However, tasks that require reasoning still require knowledge that is symbolically represented. Substantial breakthroughs can be achieved by combining deep learning with symbolic reasoning. What hinders these developments is mainly the cost of building knowledge bases rich not only in factual information, which is relatively abundant and easy to capture, but also in rules, constraints, and relationships (in summary, axioms) that make it possible to infer implicit knowledge by reasoning. The definition of the standards that collectively go under the name of “semantic Web” has provided a technological framework to produce open data as well as to define vocabularies and
ontologies to make those data interoperable. Nowadays, a huge mass of machine-readable knowledge is available on the semantic Web, which opens up enormous opportunities for research. An obvious thing to do is to analyze it and learn new knowledge from it. Potential applications range from bio-informatics to computational finance.

Objectives

The main goal of this thesis is to combine symbolic reasoning and active learning to make the automatic discovery of axioms possible, thus helping to overcome the knowledge acquisition bottleneck, while radically changing the way we look at the semantic Web: instead of postulating an a priori conceptualization of reality (i.e., an ontology) and requiring that our knowledge about facts complies with it, we propose to start from collected observations about facts and learn an ontology which is able to account for them.

Discovering new axioms from a knowledge base containing both axioms (background knowledge) and assertions (facts) may be regarded as a sort of generate and test procedure, whereas candidate axioms are generated following some heuristics and then tested to determine whether they are compatible with the facts recorded in the knowledge base and consistent with background knowledge.

The main problem is that testing a candidate axiom requires reasoning with the knowledge base plus the axiom, which can be computationally very expensive. Therefore, testing every candidate axiom would be prohibitive. A way to overcome this problem is to learn a model capable of predicting whether a candidate axiom will fit the knowledge base or not, as a surrogate for reasoning. Reasoning, however, remains an option which can be used, every once in a while, as an “oracle” to classify those candidate axioms for which the trained model has a hard time to make a reliable prediction. This looks like a perfect scenario for applying active learning. Indeed, the intuition behind active learning is that a machine learning algorithm with few labeled data can improve its result if it is allowed to choose which data to use during the learning process. The general context can be described as follows: (i) the learner has few labeled data which it uses to construct an initial model, (ii) a large set of unlabeled data is also available, (iii) an “oracle” can be asked to associate labels to some unlabeled data. The main problem is then to determine when and to which data the learner will ask a label, the aim being to revise the model in order to improve it. In this thesis, the reasoner will take the place of the “oracle” and the new axioms to be tested will take the place of the set of unlabeled data.

References

Context

The rise of deep learning has been made possible by the availability of large computing powers at affordable prices. Its performance has made it possible to solve problems that previously seemed out of reach, especially in the field of perception, such as computer vision and the analysis of natural language texts. However, tasks that require reasoning still require knowledge that is symbolically represented. Substantial breakthroughs can be achieved by combining deep learning with symbolic reasoning. What hinders these developments is mainly the cost of building knowledge bases rich not only in factual information, which is relatively abundant and easy to capture, but also in rules, constraints, and relationships (in summary, axioms) that make it possible to infer implicit knowledge by reasoning.

The definition of the standards that collectively go under the name of “semantic Web” has provided a technological framework to produce open data as well as to define vocabularies and ontologies to make those data interoperable. Nowadays, a huge mass of machine-readable knowledge is available on the semantic Web, which opens up enormous opportunities for research. An obvious thing to do is to analyze it and learn new knowledge from it. Potential applications range from bio-informatics to computational finance.

Objectives

The main goal of this thesis is to combine symbolic reasoning and active learning to make the automatic discovery of axioms possible, thus helping to overcome the knowledge acquisition bottleneck, while radically changing the way we look at the semantic Web: instead of postulating
an a priori conceptualization of reality (i.e., an ontology) and requiring that our knowledge about facts complies with it, we propose to start from collected observations about facts and learn an ontology which is able to account for them.

Discovering new axioms from a knowledge base containing both axioms (background knowledge) and assertions (facts) may be regarded as a sort of generate and test procedure, whereas candidate axioms are generated following some heuristics and then tested to determine whether they are compatible with the facts recorded in the knowledge base and consistent with background knowledge.

The main problem is that testing a candidate axiom requires reasoning with the knowledge base plus the axiom, which can be computationally very expensive. Therefore, testing every candidate axiom would be prohibitive. A way to overcome this problem is to learn a model capable of predicting whether a candidate axiom will fit the knowledge base or not, as a surrogate for reasoning. Reasoning, however, remains an option which can be used, every once in a while, as an “oracle” to classify those candidate axioms for which the trained model has a hard time to make a reliable prediction. This looks like a perfect scenario for applying active learning. Indeed, the intuition behind active learning is that a machine learning algorithm with few labeled data can improve its result if it is allowed to choose which data to use during the learning process. The general context can be described as follows: (i) the learner has few labeled data which it uses to construct an initial model, (ii) a large set of unlabeled data is also available, (iii) an “oracle” can be asked to associate labels to some unlabeled data. The main problem is then to determine when and to which data the learner will ask a label, the aim being to revise the model in order to improve it. In this thesis, the reasoner will take the place of the “oracle” and the new axioms to be tested will take the place of the set of unlabeled data.

References

[8] Claudia d'Amato, Steffen Staab, Andrea G. B. Tettamanzi, Duc Minh Tran, Fabien L. Gandon: Ontology enrichment by discovering multi-relational association rules from ontological knowledge


URL: http://www.i3s.unice.fr/~tettaman/theses.html#activelearning