Semantics for Big Data
AAAI Technical Report FS-13-04

Web Scale Information Extraction with LODIE*

Anna Lisa Gentile and **Ziqi Zhang** and **Fabio Ciravegna**
{a.l.gentile, z.zhang, f.ciravegna}@dcs.shef.ac.uk
Department of Computer Science, The University of Sheffield, UK
Regent Court, 211 Portobello, S1 4DP, Sheffield, UK

Abstract

Information Extraction (IE) is the technique for transforming unstructured textual data into structured representation that can be understood by machines. The exponential growth of the Web generates an exceptional quantity of data for which automatic knowledge capture is essential. This work describes the methodology for Web scale Information Extraction adopted by the LODIE project (Linked Open Data Information Extraction). LODIE aims to develop Information Extraction techniques able to (i) scale at web level and (ii) adapt to user information need. The core idea behind LODIE is the usage of Linked Open Data, a very large-scale information resource, as a ground-breaking solution for IE, which provides invaluable annotated data on a growing number of domains.

Introduction

Extracting information from a gigantic data source such as the Web has been considered a major research challenge, and over the years many different approaches (Etzioni et al. 2004; Banko et al. 2007; Carlson et al. 2010; Freedman and Ramshaw 2011; Nakashole, Theobald, and Weikum 2011) have been proposed. Nevertheless, the current state of the art has mainly addressed tasks for which resources for training are available (e.g. the TAP ontology in (Etzioni et al. 2004)) or use generic patterns to extract generic facts (e.g. (Banko et al. 2007), OpenCalais.com). The limited availability of resources for training has so far prevented the study of the generalised use of large-scale resources to port to specific user information needs. This paper introduces the Linked Open Data Information Extraction (LODIE) project, which focuses on the study, implementation and evaluation of IE models and algorithms able to perform efficient user-centric Web scale learning by exploiting Linked Open Data (LOD). Linked Data is a recommended best practice for exposing, sharing, and connecting data using URIs and RDF (www.linkeddata.org). LOD is ideally suited for supporting Web scale IE adaptation because it is: (i) very large scale, (ii) constantly growing, (iii) covering multiple domains and (iv) being used to annotate a growing number of pages that can be exploited for training. Current approaches to use LOD for Web scale IE are limited in scope to recognising tables (Mulwad et al. 2010), and extraction of specific answers from large corpora (Balog and Serdyukov 2011), but a generalised approach to the use of LOD for training large scale IE is still missing. LODIE will fill this gap by studying how an imprecise, redundant and large-scale resources like LOD can be used to support Web scale user-driven IE in an effective and efficient way. The idea behind the project is to adapt IE methods to detailed user information needs in a completely automated way, with the objective of creating very large domain-dependent and task-dependent knowledge bases.

*An earlier version of this paper was presented at SWAIE 2012 workshop, http://semanticweb.cs.vu.nl/swaie2012.

Related Work

Adapting IE methods to Web scale implies dealing with two major challenges: large scale and lack of training data. Traditional IE approaches apply learning algorithms that require large amount of training data, typically created by humans. However, creating such learning resources at Web scale is infeasible in practice; meanwhile, learning from massive training datasets can be redundant and quickly become intractable (Joachims 1999).

Typical Web scale IE methods adopt a light-weight iterative learning approach, in which the amount of training data is reduced to a handful of manually created examples called "seed data". These are searched in a large corpus to create an "annotated" dataset, whereby extraction patterns are generalised using some learning algorithms. Next, the learnt extraction patterns are re-applied to the corpus to extract new instances of the target relations or classes. Mostly these methods adopt a bootstrapping pattern where the newly learnt instances are selected to seed the next round of learning. This is often accompanied by some measures for assessing the quality of the newly learnt instances in order to control noisy data. Two well-known earlier systems in this area are Snowball (Agichtein et al. 2001) and KnowItAll (Etzioni et al. 2004; Banko et al. 2007). Snowball iteratively learns new instances of a given type of relation from a large document collection, while KnowItAll learns new entities of predefined classes from the Web. Both have inspired a number of more recent studies, including StatSnowball (Zhu 2009), ExtremeExtraction (Freedman and Ramshaw 2011), NELL (Carlson et al. 2010) and PROSPERA (Nakashole, Theobald, and Weikum 2011). Some interesting directions

undertaken by these systems include exploiting background knowledge in existing knowledge bases or ontologies to infer and validate new knowledge instances, and learning from negative seed data. While these systems learn to extract predefined types of information based on (limited) training data, the TextRunner (Banko et al. 2007) system proposes the "Open Information Extraction", a new paradigm that exploits generic patterns to extract generic facts from the Web for unlimited domains without predefined interests.

The emergence of LOD has opened an opportunity to reshape Web scale IE technologies. The underlying multi-billion triple store[1] and increasing availability of LOD-based annotated webpages (e.g., RDFa) can be invaluable resources to seed learning. Researchers are starting to consider the use of LOD for Web scale information extraction. However, so far research in this direction has just taken off and the use of Linked Data is limited. Mulwad et al. (Mulwad et al. 2010) proposed a method to interpret tables based on linked data and extract new instances of relations and entities from tables. The TREC2011 evaluation on the Related Entity Finding task (Balog and Serdyukov 2011) has proposed to use LOD to support answering generic queries in large corpora. While these are relevant to our research, full user-driven complex IE task based on LOD is still to come.

LODIE will address these gaps by focussing on the following research questions: (i) How to let users define Web-IE tasks tailored to their own needs? (ii) How to automatically obtain training data (and filter noise) from the LOD? (iii) How to combine multi-strategy learning (e.g., from both structured and unstructured contents) to avoid drifting away from the learning task? (iv) How to integrate IE results with LOD?

LODIE - User-centric Web scale IE

In LODIE we propose to develop an approach to Web scale IE that enables fully automated adaptation to specific user needs. LOD will provide ontologies to formalise the user information need, and will enable seeding learning by providing instances (triples) and webpages formally annotated via RDFa or Microformats. Such background knowledge will be used to seed semi-supervised Web scale learning.

The use of an uncontrolled and constantly evolving, community provided set of independent Web resource for large-scale training is totally untapped in the current state of the art. Research has shown that the relation between the quantity of training data and learning accuracy follows a non-linear curve with diminishing returns (Thompson, Califf, and Mooney 1999). On LOD the majority of resources are created automatically by converting legacy databases with limited or no human validation, thus errors are present (Auer et al. 2009). Similarly, community-provided resources and annotations can contain errors, imprecision (Lopez et al. 2010), spam, or even deviations from standards (Halpin, Hayes, and McCusker 2010). Also, large resources can be redundant, i.e. contain a large number of instances that contribute little to the learning task, while introducing considerable overhead. Very regular annotations present very limited variability, and hence (i) high overhead for the learners (which will have to cope with thousands of examples providing little contribution) and (ii) the high risk of overfitting the model. For this reason, LODIE will put particular focus on measures and strategies to filter background knowledge to obtain noiseless and efficient learning.

The main contributions by LODIE, discussed in this paper, will be: (i) a method to formalise user requirements for Web scale IE via LOD; (ii) methods to evaluate the quality of LOD data and to select the optimal subset to seed learning and (iii) the development of efficient, iterative, semi-supervised, multi-strategy Web scale learning methods robust to noise.

LODIE - Architecture and Methodology

We define Web scale IE as a tuple: $< T, O, C, I, A >$ where: T is the formalisation of the user information needs (i.e. an IE Task); O is the set of ontologies on the LOD. C is a large corpus (typically the Web) which can be annotated already in part (C_L) with RDFa/Microformats; we refer to the unannotated part as C_U. I represents a collection of instances (knowledge base) defined according to O; I_L is a subset of I containing instances already present on the LOD; I_U is the subset of I containing all the instances generated by the IE process when the task is executed on C. A is a set of annotations and consists of two parts: A_L are found in C_L, and A_U are created by the IE process; A_U can be the final set or the intermediate sets created to re-seed learning.

User needs formalisation The first requirement for adapting Web scale IE to specific user needs is to support users in formalising their information needs in a machine understandable format. Formally we define the user needs as a function: $T = f(O) \rightarrow O_L$ identifying a view on the LOD ontologies[2] describing the information extraction task. T will be materialised in the form of an OWL ontology. We propose two ways to define T. The baseline strategy will be bottom up and will include: (i) identifying manually relevant ontologies and concepts on the LOD and (ii) manually defining a view on them. The second, more challenging strategy will be based on the formalisation of user needs by exploiting Knowledge Patterns (KP), general templates or structures used to organise knowledge. Encyclopedic Knowledge Patterns (EKP) (Nuzzolese et al. 2011), are generated exploiting statistics about usage of links from Wikipedia[3] to select which relations are the most representative for each concept. In this direction we started collecting statistical evidence on the usage of relations in Linked Data which will lead to the generation of KP based on LOD. In particular we designed a measure to identify equivalent relations in LOD, which facilitates the user to make sense of sometimes obscure relations (Zhang et al. 2013). A user interface will guide the user to define the IE task in an effective and efficient way, also exploiting generated KPs.

[1] http://www4.wiwiss.fu-berlin.de/lodcloud

[2] A view is a smaller ontology only including concepts and relations which can describe the user need.

[3] http://en.wikipedia.org

Learning seed identification and filtering A set of triples I_L relevant to the users need are identified as side effect of the definition of T: they can be retrieved from existing LOD knowledge bases associated with the types in T. We currently use search engines like Sindice[4] to identify RDFa and Microformat A_L which are associated to the types in T (if available). To these, we will add further candidates A_U identified by searching the Web for linguistic realisation of the triples I_L. These annotations together with A_L are used by the multi-strategy learning process to create new candidate annotations and instances. Before feeding the identified annotations to the learning process, they will be filtered to ensure high quality in training data. The intuition is that to obtain good quality seeds we need to obtain a good trade off between consistency and variability of examples. Generating seed data from LOD without any filtering can be suitable for certain tasks, e.g. gazetteers generation (Gentile et al. 2013), but we expect not to be effective under machine learning settings. Therefore, we are currently working on defining a measure of consistency to filter A to prevent the learning algorithm to be misled by spurious data. Our hypothesis is that good data should present consistency with respect to the learning task. We will cast filtering as a problem of detecting noise in training data (Jiang and Zhou 2004; Valizadegan and Tan 2007). We will also introduce a variability measure. The idea is to map each $a \in A$ to a feature vector and generate a clustering of t_A. The variability of the data collection t_A should reflect the number of clusters derived naturally and the distribution of members in each cluster. Intuitively, a higher number of clusters imply a higher number of groups of different examples, which ensures more extraction patterns to be learnt to ensure coverage; while even distribution of cluster members ensures the patterns can be generalised for each group. We hypothesize the variability of each $a \in A$ to be dependent on the general variability of the collection, and on their distance to the centroid of each cluster because intuitively, the closer an element is to the centroid, the more representative it is for the cluster. At the end of the process we will have selected a subset $t_{A'} \subseteq t_A \subseteq A$.

Multi-strategy Learning The seed data identified and filtered in the previous steps are submitted to a multi-strategy learning method, which is able to work in different ways according to the type of webpages the information is located in: (i) a model M_S able to extract from regular structures such as tables and lists; (ii) a model M_W wrapping very regular web sites generated by backing databases and (iii) a model M_T for information in natural language based on lexical-syntactic extraction patterns. As for extracting from regular structures, following early work by (Limaye, Sarawagi, and Chakrabarti 2010), we will adopt a strategy able to exploit the dependencies among entities expressed in one page/site to learn to extract from that page. As an example, for tables we will build a feature model based on text in each cell and its context (e.g. column label, text from cells in the same row). For learning to wrap a site given one of its pages containing a potential reference to $a_{jW} \in A$, we start from identifying pages from the same site that are on the to do list for T and contain other $a_{iW} \in A$ of compatible type W in equivalent position (i.e. same XPath), which we assume are to be wrapped. Exploiting structural patterns of web pages for Information Extraction is often referred as Wrapper Induction (WI) (Kushmerick 1997). We implemented a brute force approach for WI, generating relevant gazetteers from LOD, which proved to be effective in a controlled experiment (Gentile et al. 2013).

Finally for all other cases, we will learn shallow patterns. As opposed to approaches based on complex machine learning algorithms (e.g. random walks in (Iria, Xia, and Zhang 2007)), we will focus on lexical-syntactic shallow pattern generalization algorithms. The patterns will be generalised from the textual context of each $a \in A$ and will be based on features such as words (lexical), part of speech (syntactic) and expected semantics such as related entity classes. The patterns are then applied to other webpages to create new candidate annotations. At the end of this process, we concatenate the candidate annotations extracted by each learning strategy and create a collection of candidates $a \in A_U$. These will refer to instances already known (I_L) as well as new instances (I_U).

Publication of new triples in the LOD We will develop methods to enable the learned knowledge to be published and integrated into the LOD by exposing a SPARQL endpoint. In order to do so, the candidates A_U identified by IE will be assigned to a URI, i.e. a unique identifier. We call this step disambiguation (Rowe and Ciravegna 2010). The core of our disambiguation process will be exploiting features to obtain the optimal representation of each candidate set. We will use both co-occurrence based features (gathered from the context of occurrence of a given noun phrase) and relational features (obtained by exploring relational properties in the ontologies) (Krishnamurthy and Mitchell 2011). As scalability is a major requirement both in terms of T and C, we will explore methods with minimum requirements in computational terms such as simple feature overlapping based methods (Banerjee and Pedersen 2002) and string distance metrics (Lopez et al. 2010). We will compare their effectiveness with that of more computationally intensive machine learning methods such as HMM (Rowe and Ciravegna 2010), random walks (Iria, Xia, and Zhang 2007) etc.

Conclusion

LODIE is a project addressing complex challenges that we believe are novel and of high interest to the scientific community. It is timely because (i) for the first time in the history of IE a very large-scale information resource is available, covering a growing number of domains and (ii) of the very recent interest in the use of Linked Data for Web extraction. A number of challenges are ahead and require the use of technologies from fields such as knowledge representation and reasoning, IE and machine learning. We intend to use knowledge patterns to formalise user requirements for Web scale IE. We will develop efficient iterative semi-supervised

[4] http://sindice.com

multi-strategy Web scale learning methods robust to noise and able to avoid drifting away when re-seeding. Particular focus will be put on efficient and robust methods: we will develop and test methods to evaluate the quality of LOD data for training and to select the optimal subset to seed learning.

Acknowledgments Part of this research has been sponsored by the EPSRC funded project LODIE, EP/J019488/1.

References

Agichtein, E.; Gravano, L.; Pavel, J.; Sokolova, V.; and Voskoboynik, A. 2001. Snowball: a prototype system for extracting relations from large text collections. *SIGMOD Rec.* 30(2):612–.

Auer, S.; Dietzold, S.; Lehmann, J.; Hellmann, S.; and Aumueller, D. 2009. Triplify: light-weight linked data publication from relational databases. In *Proc. of the 18th international conference on World wide web*, WWW '09, 621–630. New York, NY, USA: ACM.

Balog, K., and Serdyukov, P. 2011. Overview of the TREC 2010 Entity Track. In *Proc. of the Nineteenth Text REtrieval Conference (TREC 2010)*. NIST.

Banerjee, S., and Pedersen, T. 2002. An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In *CICLing '02: Proc. of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, 136–145. London, UK: Springer-Verlag.

Banko, M.; Cafarella, M.; Soderland, S.; Broadhead, M.; and Etzioni, O. 2007. Open information extraction for the web. In *IJCAI'07 Proc. of the 20th international joint conference on Artifical intelligence*, 2670–2676.

Carlson, A.; Betteridge, J.; Kisiel, B.; Settles, B.; Jr, E. R. H.; and Mitchell, T. M. 2010. Toward an Architecture for Never-Ending Language Learning. In *Proc. of the Conference on Artificial Intelligence (AAAI)*, 1306–1313.

Etzioni, O.; Cafarella, M.; Downey, D.; Kok, S.; Popescu, A.-M.; Shaked, T.; Soderland, S.; Weld, D. S.; and Yates, A. 2004. Web-Scale Information Extraction in KnowItAll (Preliminary Results). In *WWW2004 Proc. of the 13th international conference on World Wide Web*, 100–110.

Freedman, M., and Ramshaw, L. 2011. Extreme extraction: machine reading in a week. In *Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1437–1446. ACL.

Gentile, A. L.; Zhang, Z.; Augenstein, I.; and Ciravegna, F. 2013. Unsupervised wrapper induction using linked data. In *Proc. of the seventh international conference on Knowledge capture*, K-CAP '13, 41–48. New York, NY, USA: ACM.

Halpin, H.; Hayes, P.; and McCusker, J. 2010. When owl: sameas isn't the same: An analysis of identity in linked data. In *Proc. of 9th International Semantic Web Conference ISWC 2010*, 305–320.

Iria, J.; Xia, L.; and Zhang, Z. 2007. Wit: web people search disambiguation using random walks. In *Proc. of the 4th International Workshop on Semantic Evaluations*, SemEval '07, 480–483. Stroudsburg, PA, USA: Association for Computational Linguistics.

Jiang, Y., and Zhou, Z.-H. 2004. Editing training data for knn classifiers with neural network ensemble. In *Advances in Neural Networks ISNN 2004*, volume 3173 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 356–361.

Joachims, T. 1999. Making large scale SVM learning practical. In *Advances in kernel methods*. MIT Press, Cambridge, MA, USA. 169–184.

Krishnamurthy, J., and Mitchell, T. 2011. Which noun phrases denote which concepts? In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 15213, 570–580.

Kushmerick, N. 1997. Wrapper Induction for information Extraction. In *IJCAI97*, 729–735.

Limaye, G.; Sarawagi, S.; and Chakrabarti, S. 2010. Annotating and Searching Web Tables Using Entities , Types and Relationships. *Proc. of the VLDB Endowment* 3(1-2):1338–1347.

Lopez, V.; Nikolov, A.; Sabou, M.; and Uren, V. 2010. Scaling up question-answering to linked data. In *Proc. of the 17th international conference on Knowledge engineering and management by the masses. EKAW10*, 193–210.

Mulwad, V.; Finin, T.; Syed, Z.; and Joshi, A. 2010. Using linked data to interpret tables. In *COLD*, volume 665 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Nakashole, N.; Theobald, M.; and Weikum, G. 2011. Scalable knowledge harvesting with high precision and high recall. In *Proc. of the fourth ACM international conference on Web search and data mining*, WSDM '11, 227–236. New York, NY, USA: ACM.

Nuzzolese, A. G.; Gangemi, A.; Presutti, V.; and Ciancarini, P. 2011. Encyclopedic knowledge patterns from wikipedia links. In *Proc. of the 10th international conference on The semantic web - Volume Part I*, ISWC'11, 520–536. Berlin, Heidelberg: Springer-Verlag.

Rowe, M., and Ciravegna, F. 2010. Disambiguating identity web references using Web 2.0 data and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web* 8(2-3):125–142.

Thompson, C. A.; Califf, M. E.; and Mooney, R. J. 1999. Active learning for natural language parsing and information extraction. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, 406–414. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Valizadegan, H., and Tan, P. 2007. Kernel Based Detection of Mislabeled Training Examples. In *Proc. of the Seventh SIAM International Conference on Data Mining*, 309–319.

Zhang, Z.; Gentile, A. L.; Augenstein, I.; Blomqvist, E.; and Ciravegna, F. 2013. Mining equivalent relations from linked data. In *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 289–293. Sofia, Bulgaria: Association for Computational Linguistics.

Zhu, J. 2009. StatSnowball : a Statistical Approach to Extracting Entity. In *WWW '09 Proc. of the 18th international conference on World wide web*, 101–110.