Linked Data as Background Knowledge for Information Extraction on the Web

Ziqi Zhang 1
Anna Lisa Gentile 2
Isabelle Augenstein 3
Department of Computer Science, University of Sheffield

Information Extraction (IE) is the technique for transforming textual data into structured representation that can be understood by machines. It is a crucial technique in enabling the Semantic Web, where increasing interest has been seen in recent years. This article reports recent progress in the LODIE project - Linked Open Data for Information Extraction, aimed at advancing Web IE to a new frontier by exploiting largely available, semantically annotated, Linked Open Data as background knowledge. We cover topics of wrapper induction, IE from semi-structured content such as tables and lists, and IE from free-text. We describe new challenges in the research and methods proposed to address them, together with summaries of recent evaluations showing encouraging results.

1. INTRODUCTION

As today's Web grows into a gigantic data source, the answers to our 'data' needs could be just a few clicks away. However, precisely locating the data and more importantly, deriving information and knowledge (i.e., 'sense-making') from large amount of data remains a major challenge, as today's Web is predominantly developed for human consumption with little consideration of machine-readability. To address this challenge, the last decade has seen increasing research on Information Extraction on the Web [Etzioni et al. 2004; Etzioni et al. 2008; Carlson et al. 2010; Freedman et al. 2011; Nakashole et al. 2011; Balog and Serdyukov 2011], the goal of which is automatically transforming the classic humans' Web into structured representation that can be understood by machines.

Set out to bring the research to a new frontier, we introduced LODIE - Linked Open Data for Information Extraction [Ciravegna et al. 2012], a Web IE project focusing on exploiting largely available Linked Open Data (LOD, or Linked Data in short) as background knowledge to build unsupervised, scalable models for IE on the Web. Linked Data[1] describes best practice for exposing, sharing, and connecting data following universal protocols including URIs and RDF, such that the data can be easily integrated and re-used[2]. The emergence of LOD has opened an opportunity to reshape Web-scale IE technologies. The underlying

[1] linkeddata.org.
[2] The term also refers to the actual data that follow the practice and that are exposed on the Web.

multi-billion triple data make an ideal resource to support Web IE because it is: (1) very large scale, (2) constantly growing, (3) covering multiple domains and (4) being used to annotate a growing number of pages that can be exploited for training IE models.

Although recent years have seen increasing interest of using LOD in some IE-related tasks [Limaye et al. 2010; Balog and Serdyukov 2011; Mulwad et al. 2013], LODIE researches the usage of LOD in various Web IE tasks and at different IE stages. Figure 1 shows an overview of the LODIE architecture, which has been described in details in [Ciravegna et al. 2012]. In general, LODIE studies (1) how to use LOD to train Web IE models, or more specifically: how to support users to define their IE tasks using as templates, hundreds of thousands of schemata used for describing LOD on the Web and how to identify from the multi-billion triple data, those specific parts to build IE models for their tasks (**define IE template and gather seed data**); (2) how to filter noisy data and select the most representative and sufficiently large subset for learning, as the largely automatically generated LOD is shown to contain errors and redundancy [Halpin and Hayes 2010; Gentile et al. 2013] (**data filtering**); (3) how to approach learning in different IE tasks depending on different levels of structurality in content and how the above two points are addressed in each specific task (**multi-strategy learning**).

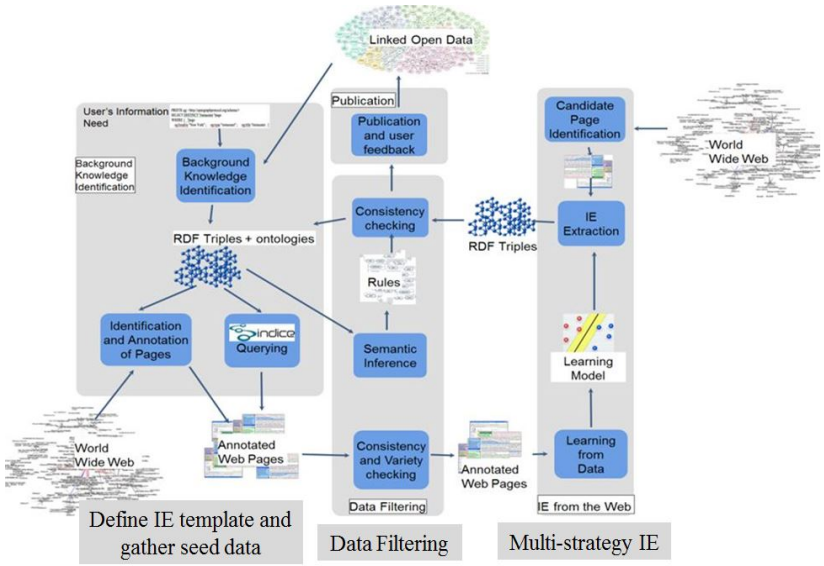

Fig. 1. Overview of the LODIE architecture.

As proof-of-concept, we have previously demonstrated methods on IE template definition and seed gathering [Blomqvist et al. 2013; Zhang et al. 2013; Zhang et al. 2013], and learning from structured Web content [Gentile et al. 2013]. In [Blomqvist et al. 2013; Zhang et al. 2013; Zhang et al. 2013], we described methods of building re-usable and extensible ontology components to reflect their usage patterns on LOD, while coping with data heterogeneity. Built on the idea of ontology design patterns [Gangemi and Presutti 2010; Nuzzolese et al. 2011], these ontological components enable users to use existing

schemata from LOD to define their IE tasks, hence avoiding the vocabulary gap [Freitas et al. 2013] and enabling seamless integration of the IE output with existing data on the Web. In [Gentile et al. 2013], we introduced a noise-robust wrapper induction method that is able to extract information from regular, re-occurring structures on a set of Web pages in an unsupervised way, benefiting from the largely available LOD datasets as seeds to train a model.

This article follows up our previous work in LODIE and provides an up-to-date overview of the recent development in the project, including methods on: data filtering to train wrapper induction more efficiently and effectively; IE from tables and lists using a bootstrapping learning approach (which partially addresses the data filtering problem) to improve efficiency and using LOD as features in learning to boost accuracy; and using triples from LOD to seed IE from free-form texts. The remainder of this article is structured as follows: Section 2 discusses related work; Section 3 to 5 summarises the above mentioned development in details. Section 6 concludes this article.

2. RELATED WORK

LODIE is most closely related to Web-scale IE methods, the most well-known of which includes Snowball [Agichtein et al. 2001], StatSnowball [Zhu et al. 2009], KnowItAll [Etzioni et al. 2004; Etzioni et al. 2008], ExtremeExtraction [Freedman et al. 2011], NELL [Carlson et al. 2010] and PROSPERA [Nakashole et al. 2011]. These methods are limited in a number of ways, which LODIE aims to tackle. First, current methods require as input a 'template' or 'schema' defining concepts and relations of interest in an IE task. However, since users may define their schemata using different vocabularies, re-using the extraction output from different Web IE systems usually requires aligning the different schemata and integrating the extracted data [Wijaya et al. 2013]. LODIE tackles this issue by enabling users to re-use existing schemata that are already used for publishing LOD on the Web.

Another problem that stems from the definition of an IE template is the provision of learning examples for every concept and relation. These are essential 'seed' data to bootstrap learning and are typically manually provided. LODIE alleviates this problem by benefiting from the gigantic, fast growing LOD not only as a source of seed data, but also a general background knowledge source providing potentially useful features for learning. IE templates defined with schemata from the LOD also make it straightforward to retrieve corresponding training data.

Further, almost all methods - with the exception of NELL - extracts information from unstructured content in free-text form only, while NELL also handles list structures. However, free-text IE is insufficient as today's Web uses extensively rich structures for content presentation. The example Web page shown in Figure 2 consists of sections displaying content following certain structured template that is typically site-specific and consistent (Part 1), content embedded in structured HTML elements such as tables (Part 2), and also unstructured paragraphs (Part 3). Note that the structures in the first two parts of the Web page carry crucial clues for extracting and interpreting the embedded content. Free-text IE techniques are often found to fail on such content [Lu et al. 2013; Mulwad et al. 2013]. LODIE aims at more comprehensive, **multi-strategy Web IE** extracting information from (1) regular, re-occurring, website-specific and consistent structures (Part 1 of Figure 2) us-

ing **wrapper induction** techniques [Kushmerick 1997; Gentile et al. 2013]; (2) structured HTML components carrying implicit semantics such as tables and lists (Part 2 of Figure 2 using **table and list interpretation** techniques [Limaye et al. 2010; Mulwad et al. 2013]; (3) unstructured text using **free-text IE** techniques [Zhou et al. 2005; Nadeau and Sekine 2007].

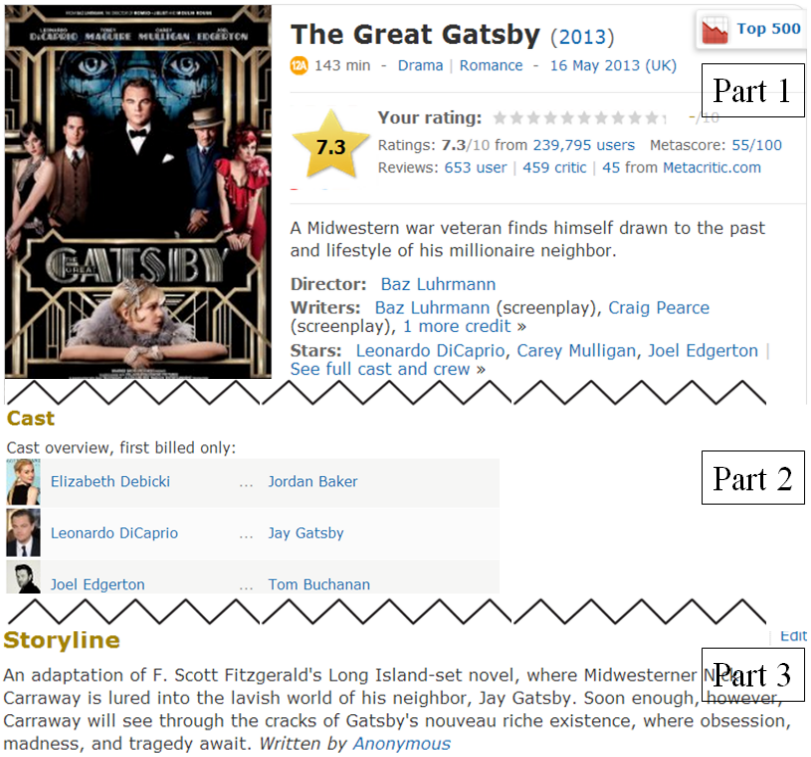

Fig. 2. Excerpts from an IMDB Web page about a film.

For each of the three component, a plethora of literature is available. While comparing and contrasting against them is not the focus of this article, the methods proposed in LODIE differentiate from the rest on two grounds: innovative usage of LOD for IE and the ability to filter noise from vast amount of data and identify most useful part of data to support learning (**data filtering**).

3. WRAPPER INDUCTION

One of the common techniques in Web IE is known as wrappers. A wrapper is generally a set of rules designed to extract data from a specific set of (semi-)structured documents that share structural similarities. They are found particularly useful for extracting information from entity-centric Web pages, usually embedded in website-specific, re-occurring structures such as Part 1 in Figure 2. We previously introduced an unsupervised wrapper induction method based on the principle of learning from large amount of (potentially)

noisy training data created by automatic dictionary-based annotation. The process begins by creating a large dictionary gazetteer for each concept of interest using triples from the LOD, then blindly annotating a corpus by matching every entry in the dictionary with text in the corpus, finally pruning the annotations and generalising extraction rules as wrappers. We showed that this simple method can obtain high accuracy with F1 between 60-100%, confirming the benefits of using very large amount of LOD in IE tasks: one does not need sohpisticated learning models in this task as the large amount of data will 'explain' themselves.

As follow-up research, we discovered that one limitation of the approach is its dependence on the quality of dictionary gazetteers. With the presence of 'noisy' entries (e.g., due to ambiguity), false positive annotations can be created, leading to false extraction patterns generalised. To illustrate, consider we are to extract film directors from Web pages like Figure 2. Since it is common to see multi-role professionals in the film industry, a person could be a director in one film but a writer, or actor/actress in another. As a result, a 'director' dictionary may contain ambiguous entries that in this very specific example, cause names of 'Writer' or 'Starring' to be annotated. Another issue is that many different dictionary entries contributed to redundant generation of the same extraction patterns, costing computation with little benefits.

To address the first issue, we proposed novel strategies to (1) reduce the number of dictionary entries likely to produce false annotations (bad seeds), and (2) detect unreliable extraction patterns independently from the seeds that contributed to the generalisation of those patterns. For (1), the principle is to discard entries that are syntactically incompatible (e.g., belong to a datatype (date, number) that is not the majority for a dictionary) or semantically ambiguous (e.g., found in different dictionaries each denoting different semantic concepts). For (2), the principle is to favour extraction patterns that are generated by a large fraction of dictionary entries, that are applicable to large proportion of the input corpus, and that extract diverse values from the corpus. To address the second issue, we proposed to use an incremental, iterative learning approach that in each turn, learns from a handful of random entries in dictionary, until no new extraction patterns can be generated. Experiments have shown that these strategies further improved learning accuracy, with an F1 of 0.85 when considering domains and websites where dictionaries can be generated using Linked Data, while reducing computation. Details of this part of research can be found in [Gentile et al. 2014]

4. TABLE AND LIST INTERPRETATION

Table and list interpretation deals with HTML table and list structures that are used to carry relational data, rather than formatting purposes. LODIE particularly focuses on table structures, which are more generic since lists can be considered as single column tables without headers. Formally, table interpretation annotates tabular data at three levels: (1) label columns with semantic concepts or properties of concepts that best describe the data contained; (2) identify the semantic relations between columns; and (3) link name mentions in table content cells with named entities in existing knowledge bases (disambiguation).

Table interpretation in LODIE addresses two major limitations in state-of-the-art. First, existing methods have predominantly adopted an *exhaustive* strategy for learning. For ex-

ample, to label the columns shown in Part 2 in Figure 2, they disambiguate every content cell in each column to derive candidate concepts. While in this very case, the original table contains over 50 rows. However, as a human reader, we can confidently label the columns (actor/actress and character played) seeing merely those three rows shown in the figure thanks to the context and our ability to infer with partial data. This ability can largely improve table interpretation efficiency. Second, state-of-the-art are almost exclusively based on two types of features: those derived from triple datasets and those derived from table components such as header text, and row content. LODIE also looks at document context that tables occur in (i.e., around and outside tables e.g., captions, page titles), which offers equally useful clues for interpretation.

To address the first limitation, LODIE proposes an incremental, bootstrapping approach that firstly learns to label table columns using partial data in the column. A method for automatically selecting the optimal part of data is also proposed. The outcome from this process can be low in accuracy, but is used as a 'stepping stone' to guide interpretation of the remaining data in the table. This generates an 'initial' interpretation of the table, which is then revised iteratively in an 'update' process that reinforces the mutual dependency between the different sub-annotation tasks (e.g., cell disambiguation generates candidate concepts for column annotation, which in return, also provides clues for disambiguation) to improve learning accuracy. To address the second limitation, LODIE proposes a generic feature model able to use various types of table context in learning. In particular, this include features from specific LOD datasets, as well as the pre-defined semantic markups within Web pages such as RDFa/Microdata[3] annotations providing important information about the Web pages and tables they contain.

Evaluation on the largest collection of table interpretation datasets known to date has shown that it significantly improves baseline exhaustive models by up to 42% in F1, while reducing CPU time by up to 29%. Details of this part of work is described in [Zhang 2014].

5. FREE-TEXT IE

Free text information extraction aims at extracting information from free-form natural language texts, such as HTML paragraph elements. In the recent progress, we focused on extracting relations from text. Our approach for relation extraction is based on the *distant supervision* paradigm [Mintz et al. 2009], which leverages on seed entities from knowledge bases to automatically annotate training data. The distant supervision assumption is that if two entities participate in a relation, any sentence that contains those two entities might express that relation.

LODIE addresses the following limitations of existing approaches: First, the distant supervision paradigm is imprecise and thus introduces noise - not all entity pairs which appear together in a sentence express the same relation. As an example, The Beatles released an album 'Let it Be' containing the song 'Let it Be'. At the stage of annotating training data, it is unclear if sentences containing both 'The Beatles' and 'Let it Be' should be used as training data for 'album' or 'song'. Using it for both will inevitably introduce noise. We propose to automatically detect and discard ambiguous seeds to reduce this

[3]E.g., with the schema.org vocabulary

noise. While there are already some approaches which attempt to reduce noise when automatically generating training data [Surdeanu et al. 2012][Roth and Klakow 2013][Riedel et al. 2010], those approaches are based on complex multi-stage machine learning models. Instead of trying to address the problem of noisy training data by using more complicated multi-stage machine learning models, we want to examine how background information extracted from LOD can be even further exploited by testing if simple statistics based on data already present in the LOD can help to filter unreliable training data. The statistics aim at assessing how likely seeds are ambiguous. Our hypothesis is that prominant values such as 'pop' for music genre are likely to be ambiguous. Evaluations on a Web crawl corpus support this hypothesis, with our best performing model achieving a precision of 0.8, whereas the same setting without filtering seeds scored a precision of 0.75 [Augenstein 2014b; 2014a].

Further, we study distant supervision in the Web context, where text content contains more noise such as spelling or grammar mistakes, and also for broader scope of domains and entity classes. Previous work however, has mainly focused on standard entity classes (e.g., PERSON, ORGANIZATION) in the news domain. Compared to the method by Mintz et al. [Mintz et al. 2009], who use the Stanford named entity tagger to recognise standard entity classes, we manage to recognise about twice as many entities using our own entity recogniser [Augenstein 2014a].

Lastly, we also find that the distant supervision assumption is quite restrictive: it requires both subject and object of a relation to be mentioned in a sentence explicitly. Using existing coreference resolution models did not significantly improve recall, however, we report positive results on a different approach for solving this by relaxing the distant supervision assumption. The new relaxed assumption is: 'if two entities participate in a relation, any paragraph that contains those two entities might express that relation, even if they are not in the same sentence, provided that another sentence in the paragraph in itself contains a relationship for the same subject.' This reduced the precision for our baseline model from 0.75 to 0.7, but resulted in three times the number of extractions [Augenstein 2014a].

6. CONCLUSION

This article discussed recent progress in LODIE, a project aimed at addressing complex Web IE tasks exploiting large amount of Linked Open Data on the Web. LODIE uses Linked Data to support various IE tasks, and at different stages of a task. So far, LODIE has developed methods and techniques of supporting users in defining their IE task templates, identifying data to seed learning, multi-strategy learning including wrapper induction, table and list interpretation and free-text IE methods, and data filtering and selection techniques specific to each learning components. An extensive set of experiments has shown that, on the one hand, the vast amount of Linked Data creates significant potential for building Web IE methods; on the other hand, being able to filter noise and identify the most optimal subset from this gigantic data source brings additional benefits to learning. Future work will target at two directions: (1) strengthen multi-strategy learning by continuing the development of each learning learning and introducing methods of integrating them to produce coherent output; (2) publishing extracted information as triples onto the LOD cloud, and address potential data integration challenges.

ACKNOWLEDGMENTS

The LODIE project (Linked Open Data Information Extraction) is funded by the Engineering and Physical Sciences Research Council, Grant Reference: 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. In *SIGMOD '01: Proceedings of the 2001 ACM SIGMOD international conference on Management of data*. ACM Press, New York, NY, USA, 612.

AUGENSTEIN, I. 2014a. Joint Information Extraction from the Web using Linked Data. *Proceedings of ISWC*, 505–512.

AUGENSTEIN, I. 2014b. Seed Selection for Distantly Supervised Web-Based Relation Extraction. *Proceedings of the COLING Workshop on Semantic Web and Information Extraction*.

BALOG, K. AND SERDYUKOV, P. 2011. Overview of the TREC 2010 Entity Track. In *Proceedings of the Nineteenth Text REtrieval Conference (TREC 2010)*. NIST.

BLOMQVIST, E., ZHANG, Z., GENTILE, A. L., AUGENSTEIN, I., AND CIRAVEGNA, F. 2013. Statistical knowledge patterns for characterizing linked data. In *Proceedings of the Workshop on Ontology and Semantic Web Patterns (4th edition) - WOP2013*. Lecture Notes in Computer Science. Springer.

CARLSON, A., BETTERIDGE, J., KISIEL, B., SETTLES, B., JR., E. H., AND MITCHELL, T. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*. AAAI Press, 1306–1313.

CIRAVEGNA, F., GENTILE, A. L., AND ZHANG, Z. 2012. Lodie: Linked open data for web-scale information extraction. In *SWAIE*, D. Maynard, M. van Erp, and B. Davis, Eds. CEUR Workshop Proceedings, vol. 925. CEUR-WS.org, 11–22.

ETZIONI, O., BANKO, M., SODERLAND, S., AND WELD, D. S. 2008. Open information extraction from the web. *Commun. ACM 51,* 12 (Dec.), 68–74.

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 *Proceedings of the 13th International Conference on World Wide Web*. WWW '04. ACM, New York, NY, USA, 100–110.

FREEDMAN, M., RAMSHAW, L., BOSCHEE, E., GABBARD, R., KRATKIEWICZ, G., WARD, N., AND WEISCHEDEL, R. 2011. Extreme extraction: Machine reading in a week. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Association for Computational Linguistics, Stroudsburg, PA, USA, 1437–1446.

FREITAS, A., OLIVEIRA, J. A. G., O'RIAIN, S., DA SILVA, J. C., AND CURRY, E. 2013. Querying linked data graphs using semantic relatedness: A vocabulary independent approach. *Data and Knowledge Engineering 88*, 126–141.

GANGEMI, A. AND PRESUTTI, V. 2010. Towards a pattern science for the semantic web. *Semant. web 1,* 1,2 (Apr.), 61–68.

GENTILE, A., ZHANG, Z., AND CIRAVEGNA, F. 2013. Web scale information extraction with lodie. In *AAAI Fall Symposium Series*. AAAI, 24–27.

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. ACM, New York, NY, USA, 41–48.

GENTILE, A. L., ZHANG, Z., AND FABIO, C. 2014. Self Training Wrapper Induction with Linked Data. In *17th International Conference on Text, Speech and Dialogue*. Springer, To appear.

HALPIN, H. AND HAYES, P. J. 2010. When owl:sameAs isnt the same: An analysis of identity links on the semantic web. In *ESWC2010*.

KUSHMERICK, N. 1997. Wrapper induction for information extraction. Ph.D. thesis. AAI9819266.

LIMAYE, G., SARAWAGI, S., AND CHAKRABARTI, S. 2010. Annotating and Searching Web Tables Using Entities , Types and Relationships. *Proceedings of the VLDB Endowment 3,* 1-2, 1338–1347.

LU, C., BING, L., LAM, W., CHAN, K., AND GU, Y. 2013. Web entity detection for semi-structured text data records with unlabeled data. *International Journal of Computational Linguistics and Applications To appear*.

MINTZ, M., BILLS, S., SNOW, R., AND JURAFSKY, D. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*. Association for Computational Linguistics, 1003–1011.

MULWAD, V., FININ, T., AND JOSHI, A. 2013. Semantic message passing for generating linked data from tables. In *International Semantic Web Conference (1)*, H. Alani, L. Kagal, A. Fokoue, P. T. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. F. Noy, C. Welty, and K. Janowicz, Eds. Lecture Notes in Computer Science, vol. 8218. Springer, 363–378.

NADEAU, D. AND SEKINE, S. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes 30,* 1 (January), 3–26. Publisher: John Benjamins Publishing Company.

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. ACM, New York, NY, USA, 227–236.

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. Springer-Verlag, Berlin, Heidelberg, 520–536.

RIEDEL, S., YAO, L., AND MCCALLUM, A. 2010. Modeling Relations and Their Mentions without Labeled Text. In *Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part III*. ECML PKDD'10. Springer-Verlag, 148–163.

ROTH, B. AND KLAKOW, D. 2013. Combining Generative and Discriminative Model Scores for Distant Supervision. In *EMNLP*. ACL, 24–29.

SURDEANU, M., TIBSHIRANI, J., NALLAPATI, R., AND MANNING, C. D. 2012. Multi-instance Multi-label Learning for Relation Extraction. In *EMNLP-CoNLL*. ACL, 455–465.

WIJAYA, D., TALUKDAR, P. P., AND MITCHELL, T. 2013. Pidgin: Ontology alignment using web text as interlingua. In *Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management*. CIKM '13. ACM, New York, NY, USA, 589–598.

ZHANG, Z. 2014. Start small, build complete: Effective and efficient semantic table interpretation using tableminer. In *Under transparent review: The Semantic Web Journal*. http://www.semantic-web-journal.net/content/start-small-build-complete-effective-and-efficient-semantic-table-interpretation-using.

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)*. Association for Computational Linguistics, Sofia, Bulgaria, 289–293.

ZHANG, Z., GENTILE, A. L., BLOMQVIST, E., AUGENSTEIN, I., AND CIRAVEGNA, F. 2013. Statistical knowledge patterns: Identifying synonymous relations in large linked datasets. In *International Semantic Web Conference (1)*, H. Alani, L. Kagal, A. Fokoue, P. T. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. F. Noy, C. Welty, and K. Janowicz, Eds. Lecture Notes in Computer Science, vol. 8218. Springer, 703–719.

ZHOU, G., SU, J., ZHANG, J., AND ZHANG, M. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. ACL '05. Association for Computational Linguistics, Stroudsburg, PA, USA, 427–434.

ZHU, J., NIE, Z., LIU, X., ZHANG, B., AND WEN, J.-R. 2009. Statsnowball: A statistical approach to extracting entity relationships. In *Proceedings of the 18th International Conference on World Wide Web*. WWW '09. ACM, New York, NY, USA, 101–110.

Author 1 and 2 are research associates in the OAK research lab in the Department of Computer Science of University of Sheffield. They specialise in research on Information Extraction, and is currently funded by EPSRC (UK) in the LODIE project, which studies methods of Information Extraction using background knowledge from Linked Data.

Author 3 is a PhD student in the OAK research lab. Her research interest is Web-scale Information Extraction and the Semantic Web.