Joint Information Extraction from the Web using Linked Data

Isabelle Augenstein

Department of Computer Science, The University of Sheffield, UK
{i.augenstein}@sheffield.ac.uk

Abstract. Almost all of the big name Web companies are currently engaged in building 'knowledge graphs' and these are showing significant results in improving search, email, calendaring, etc. Even the largest openly-accessible ones, such as Freebase and Wikidata, are far from complete, partly because new information is emerging so quickly. Most of the missing information is available on Web pages. To access that knowledge and populate knowledge bases, information extraction methods are necessitated. The bottleneck for information extraction systems is obtaining training data to learn classifiers. In this doctoral research, we investigate how existing data in knowledge bases can be used to automatically annotate training data to learn classifiers to in turn extract more data to expand knowledge bases. We discuss our hypotheses, approach, evaluation methods and present preliminary results.

1 Problem Statement

Since the emergence of the Semantic Web, many Linked datasets such as Freebase [5], Wikidata [31] and DBpedia [4] have been created, not only for research, but also commercial purposes. These have shown significant results in improving search, email, calendaring, etc. With new information emerging very quickly, cost-efficient methods for maintaining those datasets are very important. Since most of the missing or new information is available on Web pages, the cheapest method for automatically populating knowledge bases is to process those pages using information extraction (IE) methods. However, IE methods require training data to learn classifiers. Because manually creating training data is expensive and time-consuming, we propose to use *self-supervised learning* or *distant supervision*, a method proposed in recent years which utilises data already present in datasets to train classifiers [19]. Distant supervision is based on the assumption that if two entities participate in a relation, every sentence that contains those entities expresses that relation. Although distant supervision approaches are promising, they have so far ignored issues arising in the context of Web IE, specifically:
(i) *Incorrect labelling*: Distant supervision approaches automatically create training data by heuristically annotating sentences with relations between entity pairs contained in a knowledge base. This heuristic causes problems because some entity pairs are ambiguous and knowledge bases are incomplete.
(ii) *Unrecognised entities*: One subtask of relation extraction (RE) is entity recognition and classification (NERC). While existing NERC systems can be used, they are based on a restrictive set of entity types and are trained for different domains and thus often

fail to recognise entities of diverse types on heterogenous Web pages.
(iii) *Data sparsity*: Existing distant supervision approaches only learn to extract relations from text and only from sentences with contain explicit entity mentions. Since not all information on Web pages is contained in text and entities are not always refered to by their proper name, but also by using pronouns, this limits the number of extractions.
The goal of this work is to research novel methods, which do not require manually labelled training data, for Web information extracting to populate knowledge bases.

2 Relevancy

The contribution of this PhD research will be two-fold: the output of the Web information extraction can be validated manually and then used to populate knowledge bases, also, the Web information extraction system can be run as a service to do extraction on the fly for a given user query. This will be of interest for *Web companies* interested in expanding their knowledge graphs or improving results for search, *The Linked Data and Semantic Web community*, because it will allow to generate annotations and triples automatically across domains and reduce manual effort for populating ontologies, *The Natural Language Processing community*, because it will improve the state of the art in information extraction, and *the Machine Learning community*, because it will increase the real-world application and improve accessability of distant supervision.

3 Research Questions

Our overall research question is: is distant supervision a feasible approach for Web information extraction? How does it perform compared to unsupervised and semi-supervised methods? To answer this, the following research questions will be investigated:

1. **Seed Selection:** Is it possible to improve the precision and recall of distant supervision by strategically selecting training data? If so, by how much?
2. **Joint NERC and RE:** Is it possible to train an extraction model for joint distantly supervised entity classification and relation extraction? Will this achieve a higher precision and recall than a pipeline model which uses a state of the art supervised NERC as input for a distantly supervised relation extractor?
3. **Joint text, list and table extraction:** Does training a joint distantly supervised model for free text, list and table extraction from Web pages achieve higher precision and recall than using a pipeline model which combines those strategies?

4 Related Work

Web IE approaches to populate knowledge bases which try to minimise manual effort either use *semi-supervised* or *unsupervised* learning. *Semi-supervised bootstrapping* approaches such as NELL [6], PROPERA [20] and BOA [11] use pre-defined natural language patterns to extract information, then iteratively learn new patterns. While they can be used for Web IE for the purpose of populating knowledge bases, they are rule-based,

not statistical approaches, and as such make hard judgements based on prominant extraction patterns instead of soft judgments based on weights for features. Extraction patterns often have a good performance on narrow domains, but are less suitable for heterogenous domains, because they are less robust to unseen information or infrequent expressions. *Open information extraction* approaches such as TextRunner [34], Reverb [10], OLLIE [16] and ClausIE [8] are unsupervised approaches, which learn relation-independent extraction patterns from text. Although those patterns can be mapped to ontologies later, this is an error-prone process. In addition, those approaches often produce uninformative or incoherent IE patterns. *Automatic ontology learning and population approaches* such as FRED [22] and LODifier [3] extract ontology schemas and information for those schemas by performing deep semantic processing using a pipeline of text processing tools. Because those tools are trained on newswire, they are not robust enough to process noisy Web pages. Existing *distant supervision* system have so far only been developed for extraction from newswire [33], Wikipedia [19] or biomedical data [24]. They therefore fail to address issues arising when processing heterogenous Web text, such as dealing with grammar and spelling mistakes and recognising entities of diverse types.

While there is no system that incorporates all of the aspects discussed in Section 3, there are approaches which address the three invidual aspects.

Seed Selection: A few strategies for seed selection for distant supervision have already been investigated: at-least-one models [13][29][23][33][17], hierarchical topic models [1][25], pattern correlations [30], and an information retrieval approach [32]. At-least-one models assume that "if two entities particpate in a relation, at least one sentence that mentions these two entities might express that relation". While positive results have been reported for those models, Riedel et al. [23] argue that they are challenging to train because they are quite complex. Hierarchical topic model approaches group relations by assuming that the context of a relation is either specific for the pair of entities, the relation, or neither. Min et al. [17] propose a hierarchical model to only learn from positive examples to address the problem of incomplete negative training data. Takamatsu et al. [30] use a probabilistic graphical model to group extraction patterns. Xu et al. [32] propose a two-step model based on the idea of pseudo-relevance feedback. Our approach to filter unreliable training data is based on a different assumption: 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 data already present in the knowledge base can be even further exploited for simple statistical methods.

Joint NERC and RE: While RE approaches typically use a separate NERC, previous works have shown that applying text processing models in a pipeline fashion causes errors made by one component to be propagated to the next one, which has a significant impact on precision [26][27][14][15]. Approaches such as Integer Linear Programming [26][27] and Markov Logic Networks [9] have been proposed to solve both tasks at the same time. Existing distant supervision systems are based on pipeline models using supervised NERC models. This is partly because, in order to jointly solve both tasks in one fully distantly supervised model, the NERC has to be distantly supervised too.

Joint text, list and table extraction: Most existing Web extraction approaches focus on either text, list or table extraction. There are a few approaches which combine those [28][6][7][12][21][18], but they do so by using separate classifiers for the different

tasks, even apply them to different corpora, then combine the results. We argue that by considering text, tables and lists in isolation, important information gets lost. We want to research how solving those tasks at the same time and also making use of Web page-level features could improve the precision and recall of Web information extraction systems.

5 Hypotheses

Our research hypotheses are as follows:

1. **Seed Selection:** Removing ambiguous training examples, as well as possible false positives using statistical methods will help to improve the precision of distant supervision approaches.
2. **Joint NERC and RE:** State of the art supervised NERCs are trained for the news domain and will therefore have relatively low precision and recall on Web pages. Distantly supervised NERCs will perform better on Web pages than supervised NERCs trained on the news domain. Joint NE and relation extraction models will achieve a higher precision than pipeline models. Using fine-grained ontology-based NE classes instead of broad NE classes will lead to a higher RE precision.
3. **Joint text, list and table extraction:** A distant supervision model trained on combined feature vectors for text, list and table features will perform better than three separate models. Semi-structured (list and table) extractors have a substantially higher precision than unstructured (free text) extractors, which can be exploited by giving a higher weight to semi-structured features. Existing semi-structured extractors only consider lists and tables in isolation. Using the local and global context of lists and tables on Web pages as features will improve the precision of semi-structured extractors.

6 Approach

We develop a distantly supervised IE system in order to test our different hypotheses. A high-level overview of our approach in provided in Figure 1. Our approach consists the following components: a user integration component, a seed selection component, a feature extraction component and a multi-task learning component.
User Integration: The user can select what information about an entity of a specific class to extract, e.g. all members and albums of the band “The Beatles”. For evaluation purposes, those user queries will be generated automatically. Web pages for the user query are then retrieved. After the information is extracted, it is presented to the user.
Seed selection: The seed selection component decides which of the triples in the knowledge base to use for automatically annotating training data. We use several statistical measures to choose positive and negative training data. Our main idea is to select triples which have a relatively low ambiguity and are therefore very specific to the relation. As an example, The Beatles released an album called “Let it Be”, which also contains the track “Let it Be”. If a sentence contained both “The Beatles” and “Let it Be” it would be unclear if the sentence represents the relation “has album” or “has track”. We would therefore discard “Let it Be” as training data because it has a high ambiguity.

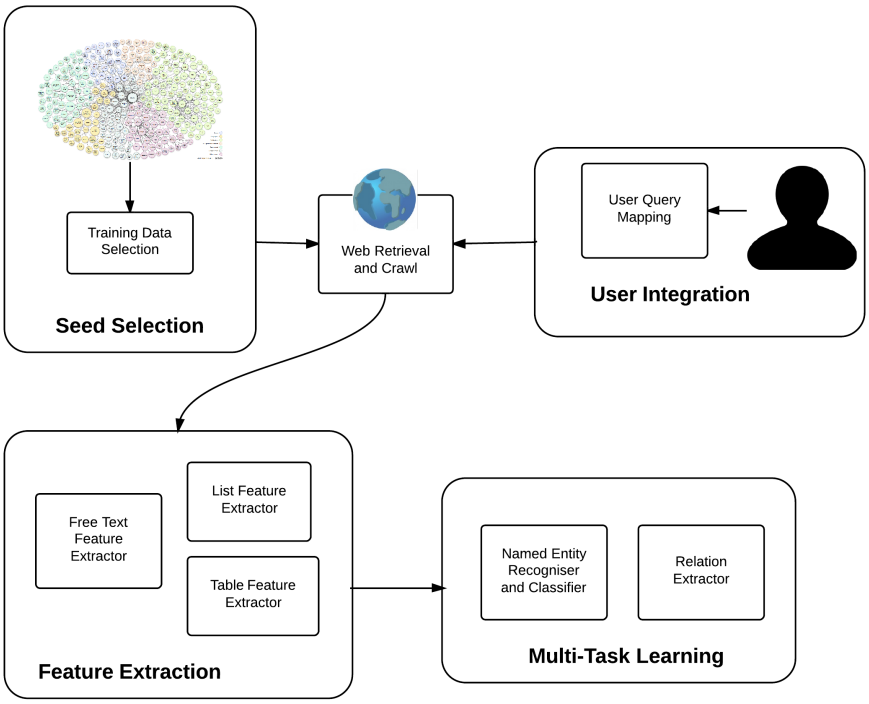

Fig. 1. Architecture of the Joint Web Extraction approach

Further, if two entities appear together in a sentence which are not related in the knowledge base, we use them as negative training data. Because knowledge bases are prone to be imcomplete, this assumption leads to further noise. Our approach is to devise statistical measures based on background data to determine how likely it is for two unrelated entities to be false positives, i.e. missing from the knowledge base. We then only select pairs of entities as negative training data which are likely to be true negatives.

Feature extraction: The feature extraction component extracts features for entities and relations in text, lists and tables on Web pages. Instead of merely considering those sources in isolation, we use the local and global context of lists and tables on Web pages as features. The local context would for example be the text appearing immediately before a list or table, whereas the global context would be the title of a Web page or words appearing on the Web page as a whole. We will also use annotations on Web pages as features, for example formatting information, which might indicate entity boundaries, and existing semantic annotations, which might help to extract relations.

Multi-task learning: The multi-task learning component learns to extract entities and relations at the same time. Further, selectional restrictions for the subject and object of a relation obtained from the knowledge base are enforced, e.g. the subject of "has album" has to be a musical artist and the object has to be an album. We use an existing

multi-task model for this. Our main contribution is to devise a distantly supervised NERC and to test if multi-task models also perform better than pipeline models if they are distantly supervised. So far, joint models have only been researched for supervised approaches. Because distantly supervised models are trained using noisy data, they are more difficult to learn than supervised ones, and using them to learn difficult models does not neccessarily improve the precision of extraction results.

7 Evaluation plan

Since the task we are working on is fairly novel, there is no existing benchmark. Existing distant supervision systems [19][23] adopt the following evaluation procedure: 1) automatic evaluation: they automatically annotate a corpus, then split it equally for training and testing; 2) manual annotation: the highest-ranked results are annotated manually; 3) precision and relative recall is computed for the automatic evaluation results and precision for the manual evaluation results. While automatic evaluation does not provide exact results, it is helpful for feature tuning. We also follow this evaluation procedure, then re-implement and compare results of existing systems on our corpus. For this purpose, we collect our own Web corpus by using focussed Web queries which contain the name of the class we are interested in (e.g. 'Book'), the name of an entity of that class, and the name of a property. Although we might be able to re-use existing Web corpora (e.g. ClueWeb [1]), it would be more difficult to locate relevant training data, and since they do not contain annotations relevant to our task, there would be no added benefit.

8 Preliminary Results

We have already performed experiments to test our first research question, results are partly documented in Augenstein [2]. We find that statistical methods for discarding highly ambiguous seeds can result in an error reduction of about 35%, however, detecting and discarding unreliable negative training data is a lot more challenging. Further, using existing NERC tools results in a low recall, since Stanford NERC is trained on a different sort of text, news, and uses standard NE classes (person, location, organisation, mixed). It often fails to recognise entities which should fall into the 'mixed' class, such as 'track' or 'album'. Recall improves significantly when using our own NER in addition to Stanford NERC, which indicates that further experiments on distantly supervised NERC for RE might be useful. Lastly, we find that the distant supervision assumption is quite restrictive: It requires both subject and object of a relation to be mentioned in the same sentence explicitly. Using existing coreference resolution tools does not significantly improve recall, however, we find that performing RE across sentence boundaries by relaxing the distant supervision assumption results in three times the number of extractions.

9 Reflections

Our approach aims at learning to extract information from the Web in a novel way. Traditional approaches use supervised learning to train models for extracting entities

[1] http://lemurproject.org/clueweb12/

and relations. Our approach is based on distant supervision, a method that has gained popularity recently, which leverages on Linked Data to train extraction models and does not require manually labelled training data. Distant supervision has so far only been applied to RE from text and not been used for Web information extraction. We argue that Web IE is more challenging than the well-researched task of IE from newswire since Web content often contains noise such as spelling or grammar mistakes, but it is also more useful for gathering information, since most content is available on the Web. While existing approaches focus on standard NE classes, we argue that it would be beneficial to extend this approach to NE classes of any domain and for this purpose investigate distantly supervised NERC for RE. Preliminary results suggest this would substantially increase the number of extractions [2]. We further find that filtering unreliable training data using statistical methods results in an error reduction of about 35% [2].

Acknowledgements

We thank Fabio Ciravegna and Diana Maynard for helping to develop this research plan, Ruben Verborgh and Tom De Nies for their writing tips, as well as the anonymous reviewers for their valuable feedback. This research is partly supported by the EPSRC funded project LODIE: Linked Open Data for Information Extraction, EP/J019488/1.

References

1. Alfonseca, E., Filippova, K., Delort, J.Y., Garrido, G.: Pattern Learning for Relation Extraction with a Hierarchical Topic Model. In: Proceedings of ACL. Jeju, South Korea (2012)
2. Augenstein, I.: Seed Selection for Distantly Supervised Web-Based Relation Extraction. In: Proceedings of SWAIE (2014)
3. Augenstein, I., Padó, S., Rudolph, S.: LODifier: Generating Linked Data from Unstructured Text. In: Proceedings of ESWC. pp. 210–224 (2012)
4. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia-A crystallization point for the Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web 7(3), 154–165 (2009)
5. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: A Collaboratively Created Graph Database For Structuring Human Knowledge. In: Proceedings of ACM SIGMOD. pp. 1247–1250 (2008)
6. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E.R., Mitchell, T.M.: Toward an Architecture for Never-Ending Language Learning. In: In AAAI (2010)
7. Carlson, A., Betteridge, J., Wang, R.C., Jr., E.R.H., Mitchell, T.M.: Coupled Semi-Supervised Learning for Information Extraction. In: Proceedings of WSDM (2010)
8. Del Corro, L., Gemulla, R.: ClausIE: Clause-Based Open Information Extraction. In: Proceedings of the 22nd International Conference on World Wide Web. pp. 355–366 (2013)
9. Domingos, P., Kok, S., Lowd, D., Poon, H., Richardson, M., Singla, P.: Markov logic. In: Probabilistic Inductive Logic Programming. pp. 92–117 (2008)
10. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: Proceedings of EMNLP. pp. 1535–1545 (2011)
11. Gerber, D., Ngomo, A.C.N., Gerber, D., Ngomo, A.C.N., Unger, C., Bühmann, L., Lehmann, J., Ngomo, A.C.N., Gerber, D., Cimiano, P.: Extracting Multilingual Natural-Language Patterns for RDF Predicates. In: Proceedings of EKAW. pp. 87–96 (2012)

12. Govindaraju, V., Zhang, C., Ré, C.: Understanding Tables in Context Using Standard NLP Toolkits. In: Proceedings of ACL (2013)
13. Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L.S., Weld, D.S.: Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations. In: Proceedings of ACL. pp. 541–550 (2011)
14. Kate, R.J., Mooney, R.J.: Joint Entity and Relation Extraction using Card-Pyramid Parsing. In: Proceedings of CoNLL. pp. 203–212 (2010)
15. Li, Q., Heng, J.: Incremental Joint Extraction of Entity Mentions and Relations. In: Proceedings of ACL (2014)
16. Mausam, Schmitz, M., Soderland, S., Bart, R., Etzioni, O.: Open Language Learning for Information Extraction. In: Proceedings of EMNLP-CoNLL. pp. 523–534 (2012)
17. Min, B., Grishman, R., Wan, L., Wang, C., Gondek, D.: Distant Supervision for Relation Extraction with an Incomplete Knowledge Base. In: Proceedings of HLT-NAACL. pp. 777–782 (2013)
18. Min, B., Shi, S., Grishman, R., Lin, C.Y.: Ensemble Semantics for Large-scale Unsupervised Relation Extraction. In: EMNLP-CoNLL. pp. 1027–1037. ACL (2012)
19. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of ACL. vol. 2, pp. 1003–1011 (2009)
20. Nakashole, U., Theobald, M., Weikum, G.: Scalable Knowledge Harvesting with High Precision and High Recall. In: Proceedings of WSDM. pp. 227–236 (2011)
21. Pennacchiotti, M., Pantel, P.: Entity Extraction via Ensemble Semantics. In: Proceedings of EMNLP. pp. 238–247 (2009)
22. Presutti, V., Draicchio, F., Gangemi, A.: Knowledge Extraction Based on Discourse Representation Theory and Linguistic Frames. In: Proceedings of EKAW. pp. 114–129 (2012)
23. Riedel, S., Yao, L., McCallum, A.: Modeling Relations and Their Mentions without Labeled Text. In: Proceedings of ECML-PKDD. pp. 148–163 (2010)
24. Roller, R., Stevenson, M.: Self-Supervised Relation Extraction using UMLS. In: Proceedings of CLEF (2014)
25. Roth, B., Klakow, D.: Combining Generative and Discriminative Model Scores for Distant Supervision. In: Proceedings of ACL-EMNLP. pp. 24–29 (2013)
26. Roth, D., tau Yih, W.: A Linear Programming Formulation for Global Inference in Natural Language Tasks. In: Proceedings of CoNLL. pp. 1–8 (2004)
27. Roth, D., Yih, W.t.: Global Inference for Entity and Relation Identification via a Linear Programming Formulation. Introduction to Statistical Relational Learning pp. 553–580 (2007)
28. Shinzato, K., Torisawa, K.: Acquiring Hyponymy Relations from Web Documents. In: HLT-NAACL. pp. 73–80 (2004)
29. Surdeanu, M., Tibshirani, J., Nallapati, R., Manning, C.D.: Multi-instance Multi-label Learning for Relation Extraction. In: Proceedings of EMNLP-CoNLL. pp. 455–465 (2012)
30. Takamatsu, S., Sato, I., Nakagawa, H.: Reducing Wrong Labels in Distant Supervision for Relation Extraction. In: Proceedings of ACL. pp. 721–729 (2012)
31. Vrandečić, D., Krötzsch, M.: Wikidata: A Free Collaborative Knowledge Base. Commun. ACM (2014), to appear
32. Xu, W., Hoffmann, R., Zhao, L., Grishman, R.: Filling Knowledge Base Gaps for Distant Supervision of Relation Extraction. In: Proceedings of ACL. pp. 665–670 (2013)
33. Yao, L., Riedel, S., McCallum, A.: Collective Cross-document Relation Extraction Without Labelled Data. In: Proceedings of EMNLP. pp. 1013–1023 (2010)
34. Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., Soderland, S.: TextRunner: Open Information Extraction on the Web. In: Proceedings of HLT-NAACL: Demonstrations. pp. 25–26 (2007)