Out of Many, One: Unifying Web-Extracted Knowledge Bases

Mathias Niepert
Computer Science and Engineering
University of Washington
Seattle, WA, 98195

Sameer Singh
Computer Science and Engineering
University of Washington
Seattle, WA, 98195

Extracting knowledge from large text corpora and the world wide web is an important problem in artificial intelligence. Arguably, the majority of the world's knowledge is contained in natural language text and as such needs to be brought into structured form to be accessible for automated reasoning. There are numerous information extraction (IE) projects that address this problem, such as YAGO, Freebase, and OpenIE [10, 11]. Each of these projects has its unique strengths and weaknesses. For instance, projects unconstrained by an ontology provide more coverage, but suffer from noise and ambiguities of the extracted facts. If these projects are categorized along dimensions such as extraction types, temporal and geographical attributes, events coverage, and schema language, one realizes that the projects would be highly complementary if their knowledge was integrated.

In this paper, we pose the problem of unifying web-extracted knowledge bases into a consistent global model as a research challenge for the academic community. We believe that a unified model should (a) facilitate statistical-relational inference across all the KBs; (b) make explicit the approximate alignments between relations, attributes, types, entities, and their mentions; and (c) feature a powerful query language that allows one to query the model via mentions. Such a system will be able to combine the advantages of using ontology-free, flexible but difficult to query KBs such as OpenIE and NELL with rigid but precisely-defined fixed-schema KBs such as YAGO, DBPedia, and Freebase. We state the problem precisely, position it in the context of existing work, discuss some challenges, and conjecture the first steps towards a solution.

A large and growing body of literature has investigated problems related to aligning KBs. Downey et al. [8] consider the task of integrating automatically extracted KBs, and propose an interesting approach of using natural language as the interface for integrating and evaluating extracted KBs, however considerable challenges in language generation and entailment inference need to be addressed before high-quality evaluation can be performed using natural language. Aligning multiple schematized KBs has also received significant attention, for example Wijaya et al. [24] and Suchanek et al. [21]. Similarly, Mahdisoltani et al. [14] introduces multilingual YAGO, which aligns the attributes and relations across languages. Linking schema-free triplets to a schematized KB is similar to information extraction, and a number of approaches have been proposed for it. Lin et al. [13] identify and link noun phrases to Wikipedia pages in order to unify OpenIE with Wikipedia-based KBs, such as DBPedia and YAGO. Low-dimensional embedding of schema-free text (surface forms) and the contents of a canonicalized KB (Freebase) have been used for aligning relations [19] and entity types [25] independently. All of these approaches are restricted by their focus on aligning a single aspect of the KB, and either ignore other aspects (relations in Lin et al. [13]) or assume they are solved (entity linking in Riedel et al. [19], Yao et al. [25]). Preliminary work on epistemological databases [23] provides an intuitive framework for integrating multiple sources of evidence, however it is not clear how to jointly model all the various aspects of the KBs, and whether inference will scale to such complex, joint models over large KBs. Since these approaches address crucial subcomponents of our proposed problem, it will be interesting to investigate how they can be extended to create a joint, unified KB from multiple, noisy KBs.

While these above approach have made considerable steps towards unifying extracted KBs, they each address a different subset of the problem; the goal here is more ambitious. We argue that to address the proposed problem, the solution would consist of a joint probabilistic model that (a) aligns

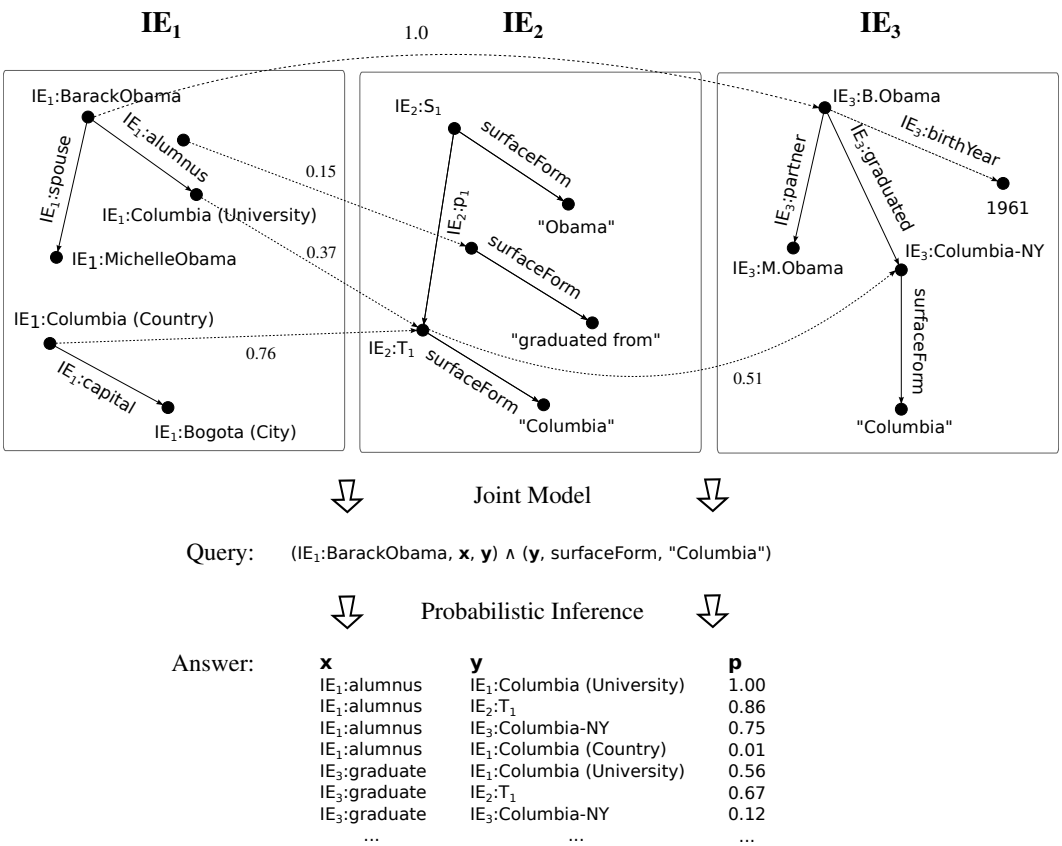

Figure 1: We propose to unify several information extraction projects into a consistent global model, integrating individuals, types, relations, and attributes. Here, the links between different KBs are equivalence correspondences. A-priori confidences for these correspondences could be based on some distributional representation such as a joint low-dimensional embedding of the KBs. Queries are either maximum a-posteriori queries or (conditional) probability queries in the form of unions of conjunctive queries [22]. The joint probabilistic model is based on a possible world semantics where, intuitively, every possible world corresponds to one deterministic alignment of the IE projects.

relations, attributes, types, and entities probabilistically, (b) utilizes distributions of types, relations, and attributes for an accurate alignment, and (c) features a rich query language supporting downstream tasks such as schema-free question answering. As the number of automatically constructed KBs grows, each providing its own set of advantages, there is a significant need for approaches that jointly integrate the entities, relation, types, and attributes of these KBs to build a coherent, queryable, joint representation.

1 Problem Formulation: KB Unification

Before we introduce the problem of unifying web-extracted KBs, we have to define what a knowledge base is. A knowledge base represents *individuals* (Albert Einstein, Mount Everest); *types* (`Actors`, `Scientists`, `Mountains`); *relations* (`studentOf`, `knows`); and *attributes* (`age`, `geoLocation`, `surfaceForm`). We use the word *object* to refer to any individual, type, relation, or attribute. Relations are sets of pairs of individuals whereas attributes are sets of entity-value pairs, relating entities to the respective attribute domain. For instance, the relation `studentOf` relates students and their teachers and the attribute `elevation` relates places and their altitude in meters. A knowledge base is a collection of triples (s, p, t) consisting of a *subject*, *predicate*, and *target*. *Subjects* and *targets* are objects of the knowledge base. In addition, *targets* can also be values of an attribute domain such as the reals. A *Predicate* is either a relation (`studentOf`, `knows`) or an attribute (`age`, `heights`, `elevation`). There are some reserved relations such as `type`, `subTypeOf`, and

subRelationOf, modeling types as well as type and relation hierarchies. The logical semantics of the KB is based on the RDF and RDF(S) semantics which supports several query types such as conjunctive queries. Figure 1 depicts fragments of three web-extracted KBs with prior similarity scores between entities. Further, each triplet may also have a confidence score associated with it; for triplets without a confidence score, we assume a KB-specific prior can be used to assign a default confidence.

Note that extractions from IE projects that do not have neither canonical identifiers nor an underlying ontology can be also be directly represented in the above formalism. One simply has to translate the triples, introducing canonical identifiers for the latent entities. For instance, the Open IE triple $(\texttt{Germany}, \texttt{plays}, \texttt{today})$ is translated to the four triples $(\texttt{s1}, \texttt{surfaceForm}, \text{“Germany”}), (\texttt{p1}, \texttt{surfaceForm}, \text{“plays”}), (\texttt{t1}, \texttt{surfaceForm}, \text{“today”})$, and $(\texttt{s1}, \texttt{p1}, \texttt{t1})$, introducing $\texttt{s1}, \texttt{p1}$, and $\texttt{t1}$ as additional identifiers.

The semantics of the unified joint knowledge base is based on the possible world semantics that is commonly used in statistical relational formalism such as probabilistic databases [22] and Markov logic networks [7]. The strong assumption of independence between facts in probabilistic databases, however, make it difficult to model the complex dependencies required for the task of knowledge base unification. For instance, the probabilities of alignments between entities and types ought to have an influence on the alignments between relations and vice versa. Modeling these dependencies is also important as we strive for *complex* alignments that go beyond simple equivalence correspondences. On the other hand, while existing statistical relational languages such as Markov logic [18] and ProbLog [17] could represent most of these more complex dependencies, the inference problem in these models is intractable and existing algorithms do not scale to large problem instances. Hence, we believe that novel statistical relational formalisms are needed that strike the right balance between expressiveness and tractability. With particular applications such as open-domain question answering, entity linking, and link prediction in mind, we believe that unions of conjunctive queries [22] (UCQ) comprise the appropriate query formalism under the possible worlds semantics. Numerous natural language queries can be translated to unions of conjunctive queries. Moreover, the tractability of UCQs is well understood for probabilistic databases.

A typical conjunctive query is depicted in Figure 1. It asks for all the relationships the entity BarackObama in KB IE_1 could have with entities that have a surface form “Columbia.” As an additional example, consider again the open-domain triple $(\texttt{Germany}, \texttt{plays}, \texttt{today})$. The four triples $(\texttt{s1}, \texttt{surfaceForm}, \text{“Germany”}), (\texttt{p1}, \texttt{surfaceForm}, \text{“plays”}), (\texttt{t1}, \texttt{surfaceForm}, \text{“today”})$, and $(\texttt{s1}, \texttt{p1}, \texttt{t1})$ in the proposed representation with $\texttt{s1}, \texttt{p1}$, and $\texttt{t1}$ as variables, can directly be used as a conjunctive query that retrieves all possible groundings of the triple to other canonicalized KBs, together with their probabilities.

The problem of web-extracted KB unification can now be formally posed as follows. Given two or more IE projects and their KBs as a set of triples of the above form, unify the KBs into a global joint probabilistic model, aligning entities, relations, types, and attributes, and allowing the applications and users to query the unified representation using unions of conjunctive queries.

2 Challenges

The problem of unifying web-extracted KBs is more challenging than the common problem of aligning knowledge bases. Knowledge bases are usually assumed to have canonical identifiers and structure-providing ontologies. For the problem of unifying web-extracted KBs, on the other hand, one cannot make these assumptions. For instance, triples extracted by open IE systems are populated with mentions only and are highly noisy. Moreover, mentions occurring in an open IE based KB often have no corresponding canonical entity in other KBs. Most existing ontology matching systems assume both the facts in the KBs to be true and a one-to-one alignment, and perform poorly when these properties are not satisfied. The alignment of types and relations across web-extracted KBs is especially challenging. Since exact equivalence links between types and relations are unlikely to exist, a unifying model needs to introduce and infer various kinds of links such as subsumption between types and relations. Probabilistic links, that is links holding only with a particular probability, are needed to facilitate a global model incorporating multiple heterogeneous KBs. Most existing ontology alignment algorithms compute only non-probabilistic alignments. Finally, the computational complexity of performing joint inference over multiple very

large knowledge bases is a major challenge. If we are to use a probabilistic joint model, we have to develop tractable formalisms that scale to very large knowledge bases.

3 Research Directions

Learning a joint model that unifies several IE projects would be extremely beneficial for several NLP tasks such as entity linking, co-reference resolution, knowledge base population, and question answering. This suggests that future research should address the problems associated with this task, and we outline some of the directions we believe to be particularly promising.

3.1 Combining Distributional and Logical Semantics

There has been a recent interest in computing distributional representations such as low-dimensional embeddings to learn similarities between KB entities (both entities and relations) [3, 5, 19, 4]. Similarities computed based on distributional representations have also been combined with statistical relational languages [2]. We believe that finding highly tractable methods that combine distributional a-priori confidences for links with statistical-relational formalism are promising directions. Low-dimensional embeddings of entities, concepts, relations, and attributes are a powerful method to compute similarities and capture the distributional semantics between these entities, while also providing efficient inference through linear algebraic computations. However, it is challenging to model more complex logical rules and the KBs ontologies with these approaches. Hence, novel combinations of distributional and logical semantics are needed to approach the problem of unifying IE projects into a tractable joint model, such as by Rocktaschel et al. [20].

3.2 Unlinkable Entities, Types, and Relations

The major motivation for a unified joint model is the integration of IE projects located on a spectrum between two extremes: unstructured extractions such as OpenIE and structured extractions such as DBpedia. Due to the much broader coverage of extractions not confined to a preexisting schema, there will be many cases where entities, types, and relations of unstructured projects cannot be aligned with those of more structured ones. Recent work on the problem of unlinkable entities and types [13] has introduced the "unlinkable noun phrase problem": Given an unlinkable noun phrase, determine if it is an entity and its semantic types. We believe that the problem of detecting unlinkable objects is important as it points to missing entities, types, and relations in other KBs. Future research should focus on methods that propose novel relations and types which are either novel or refine existing types and relations.

3.3 Numerical Attributes

Embedding approaches to entity linking and relation extraction are capable of taking distributions over types, relations, and mentions into account. However, none of the existing approaches incorporate distributions over numerical attribute values which are abundant in most knowledge bases. For instance, in DBpedia, YAGO, and Freebase there are numerous numerical attributes expressing age, elevation, geographical coordinates, temporal information, and so on. Similar to collective numerical attributes estimation in Bakalov et al. [1], future approaches to unifying IE projects must incorporate the numerical attributes by learning joint distributions over attribute values, entities, types, and relations. For instance, consider the year of birth attribute present in most KBs. We should be able to learn that if two entities are in a `studentOf` relation, the difference of their birth years follows a particular distribution. Now, when we estimate links in a unified representation, we should assign a very low probability to links that result in a `studentOf` relationship between two entities whose birth years are far apart. Moreover, spatial and temporal information about places and events is extremely useful and should be exploited in the unified model.

3.4 Joint Inference Complexity

There are several promising tractable models and inference algorithms which can be extended for this particular task. For instance, random walk type algorithms in large web-extracted knowledge bases have shown remarkable performance [12]. Moreover, there are tractable variants of otherwise intractable statistical relational models that might be suitable for the task at hand. Examples are probabilistic databases [22] and tractable statistical relational languages [6, 15]. Since inference algorithms for statistical relational formalisms such as Markov logic do not scale, or scale only under strong symmetry assumptions [16], possible research should be concerned with identifying particular language restrictions that facilitate efficient inference and with approximate inference suitable for the large-scale inference problems at hand.

4 Evaluation – Datasets and Metrics

There are links between most entities of Wikipedia-based IE projects. For instance, there are thousands of links between FREEBASE and DBPEDIA. These can be leveraged as training data. Finding and evaluating alignments between Open IE and more structured, ontology-based extractions is more challenging. With the help of researchers at the University of Mannheim, we have begun to assemble a gold standard data set for evaluation purposes [9]. We currently have alignments of 1200 NELL triples with DBPEDIA entities. These alignments, however, only relate subjects and targets of the NELL triples to entities in DBPEDIA. We are also working on manually aligning OPENIE triples with DBPEDIA.

Evaluating alignments between entities is possible using metrics such as accuracy, precision, and recall. However, in many cases a crisp alignment between relations, attributes and types is not possible and one might be interested in different kinds of links between those. In these cases we propose to measure the quality of the alignment using distance measures between probability distributions such as the Kullback-Leiber divergence and Hellinger distance. For instance, in order to asses the quality of a `subRelationOf` alignment between the relations $\mathtt{r}$ and $\mathtt{r}'$ of two different KBs, we first compute a "gold standard" probability according to the given labeled data as

$$P[(\mathtt{r}, \mathtt{subRelationOf}, \mathtt{r}')] = \frac{|\{(\mathtt{s},\mathtt{t}) \mid (\mathtt{s},\mathtt{r},\mathtt{t}) \wedge (\mathtt{s},\mathtt{r}',\mathtt{t})\}|}{|\{(\mathtt{s},\mathtt{t}) \mid (\mathtt{s},\mathtt{r}',\mathtt{t})\}|}$$

where subjects $\mathtt{s}$ and targets $\mathtt{t}$ are entities from the gold standard alignment between entities. The KL divergence between the above gold standard probability and the a-posteriori probability returned by the joint probabilistic model is then utilized to asses the quality of the relation alignment. Of course, in the rare event that a crisp gold standard for subsumptions between relations and types exists, one can perform an evaluation of systems based on ranking measures [21, 24] common in the information retrieval literature.

Since creating an annotated data set that evaluates all the important aspects of KB unification is expensive and time-consuming, it is also worthwhile to consider evaluation in terms of accuracy on downstream applications of such a unified representation, such as entity linking, question answering, coreference resolution, and knowledge-base completion.

Acknowledgements

This work was supported in part by the TerraSwarm Research Center, one of six centers supported by the STARnet phase of the Focus Center Research Program (FCRP) a Semiconductor Research Corporation program sponsored by MARCO and DARPA.

References

[1] A. Bakalov, A. Fuxman, P. P. Talukdar, and S. Chakrabarti. Scad: Collective discovery of attribute values. In *International Conference on World Wide Web (WWW)*, 2011.

[2] I. Beltagy, C. Chau, G. Boleda, D. Garrette, K. Erk, and R. Mooney. Montague meets markov: Deep semantics with probabilistic logical form. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*Sem-2013)*, pages 11–21, 2013.

[3] A. Bordes, J. Weston, R. Collobert, and Y. Bengio. Learning structured embeddings of knowledge bases. In *AAAI*, 2011.

[4] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, 2013.

[5] B. Dalvi, W. W. Cohen, and J. Callan. Collectively representing semi-structured data from the web. In *Workshop on Automated Knowledge Base Construction (AKBC)*, 2012.

[6] P. Domingos and W. A. Webb. A tractable first-order probabilistic logic. In *AAAI*, 2012.

[7] P. Domingos, D. Jain, S. Kok, D. Lowd, H. Poon, and M. Richardson. Alchemy website. http://alchemy.cs.washington.edu/, 2012. last visit: 22.11.2012.

[8] D. Downey, C. S. Bhagavatula, and A. Yates. Using natural language to integrate, evaluate, and optimize extracted knowledge bases. In *CIKM Workshop on Automated Knowledge Base Construction (AKBC)*, 2013.

[9] A. Dutta, C. Meilicke, M. Niepert, and S. P. Ponzetto. Integrating open and closed information extraction: Challenges and first steps. In *Proceedings of the NLP & DBpedia workshop*, 2013.

[10] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74, 2008.

[11] O. Etzioni, A. Fader, J. Christensen, S. Soderland, and M. Mausam. Open information extraction: the second generation. In *International joint conference on Artificial Intelligence (IJCAI)*, 2011.

[12] N. Lao, T. Mitchell, and W. W. Cohen. Random walk inference and learning in a large scale knowledge base. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2011.

[13] T. Lin, Mausam, and O. Etzioni. No noun phrase left behind: Detecting and typing unlinkable entities. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2012.

[14] F. Mahdisoltani, J. A. Biega, and F. M. Suchanek. Yago3: A knowledge base from multilingual wikipedias. In *Conference on Innovative Data Systems Research (CIDR)*, 2015.

[15] M. Niepert and P. Domingos. Tractable probabilistic knowledge bases: Wikipedia and beyond. In *Proceedings of the 4th Statistical Relational Learning Workshop (StarAI)*, 2014.

[16] M. Niepert and G. Van den Broeck. Tractability through exchangeability: A new perspective on efficient probabilistic inference. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 2467–2475, 2014.

[17] L. D. Raedt, A. Kimmig, and H. Toivonen. Problog: A probabilistic prolog and its application in link discovery. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2462–2467, 2007.

[18] M. Richardson and P. Domingos. Markov logic networks. *Machine learning*, 62(1):107–136, 2006.

[19] S. Riedel, L. Yao, A. McCallum, and B. M. Marlin. Relation extraction with matrix factorization and universal schemas. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, pages 74–84, 2013.

[20] T. Rocktaschel, S. Singh, M. Bosnjak, and S. Riedel. Low-dimensional embeddings of logic. In *ACL 2014 Workshop on Semantic Parsing (SP14)*, 2014.

[21] F. M. Suchanek, S. Abiteboul, and P. Senellart. PARIS: Probabilistic Alignment of Relations, Instances, and Schema. *Proceedings of Very Large Databases (VLDB)*, 5(3):157–168, 2011.

[22] D. Suciu, D. Olteanu, R. Christopher, and C. Koch. *Probabilistic Databases*. Morgan & Claypool Publishers, 1st edition, 2011.

[23] M. Wick. *Epistemological Databases for Probabilistic Knowledge Base Construction*. PhD thesis, University of Massachusetts, Amherst, 2014.

[24] D. Wijaya, P. P. Talukdar, and T. Mitchell. Pidgin: Ontology alignment using web text as interlingua. In *Proceedings of the Conference on Information and Knowledge Management (CIKM 2013)*, San Francisco, USA, October 2013. Association for Computing Machinery.

[25] L. Yao, S. Riedel, and A. McCallum. Universal schema for entity type prediction. In *CIKM Workshop on Automated Knowledge Base Construction (AKBC)*, 2013.