Applying Universal Schemas for Domain Specific Ontology Expansion

Paul Groth, Sujit Pal, Darin McBeath, Brad Allen and **Ron Daniel**
{p.groth, sujit.pal, d.mcbeath, b.allen, r.daniel}@elsevier.com
Elsevier Labs
1600 John F. Kennedy Boulevard, Suite 1800,
Philadelphia, PA

Abstract

Manually created large scale ontologies are useful for organizing, searching, and repurposing content ranging from scientific papers and medical guidelines to images. However, maintenance of such ontologies is expensive. In this paper, we investigate the use of universal schemas (Riedel et al., 2013) as a mechanism for ontology maintenance. We apply this approach on top of two unique data sources: 14 million full-text scientific articles and chapters, plus a 1 million concept hand-curated medical ontology. We show that using a straightforward matrix factorization algorithm one can achieve 0.7 F1 measure on a link prediction task in this environment. Link prediction results can be used to suggest new relation types and relation type synonyms coming from the literature as well as predict specific new relation instances in the ontology.

1 Introduction

Scholarly information has been a key domain of interest for the automated knowledge base construction (AKBC) community (Ororbia II et al., 2014). A range of techniques have been applied to wide variety of tasks including: the construction of pathway databases (Friedman et al., 2001), genomic knowledge extraction (Poon et al., 2014); scientific question answering (Clark et al., 2016), and scientific entity search (Sinha et al., 2015).

In this work, we focus on augmenting an existing rich domain specific ontology with additional information. We show that existing well-known unsupervised approaches can generate interesting input for domain experts in their ontology maintenance task. We apply these approaches to two unique resources:

- The full text of all 14 million documents (journal papers and book chapters) within Elsevier's ScienceDirect database. This covers over 24 major disciplines and over 2500 journals.
- The Elsevier Merged Medical Taxonomy (EMMeT) - a manually curated ontology containing nearly one million concepts, three million synonyms, more than 30 relation types, and more than three million instances of relations between those concepts.

EMMeT is used within a number of Elsevier's search engines to provide structured search results. For example, within Clinical Key (a literature search engine for clinicians), a user may search for "breast cancer" but wants to know specific treatment procedures for a particular sub type of the cancer (e.g. Malignant Neoplasm of the breast outer quadrant). EMMeT allows for the traversal from superclass to subclass to procedure to be made. However, the maintenance of EMMeT is a time consuming process requiring domain experts (e.g. trained doctors) to read the literature, update, and curate the ontology. This includes not only the addition of new relations but also revising existing relations, adding new synonyms and finding additional evidence for statements within the ontology. We also aim to show that automated approaches can provide a quick and relatively high quality entry point that can augment curated knowledge bases. This work is an initial step in applying automated knowledge base techniques in this setting.

Proceedings of AKBC 2016, pages 81–85,
San Diego, California, June 12-17, 2016.

An important aspect of this work is that it shows that a rather straightforward implementation of the universal schema approach (Riedel et al., 2013) to relation extraction can be effective in a domain specific setting. Prior work has used universal schemas for more general encyclopedic knowledge bases such as Freebase or TAC KBP datasets. Here, we apply it to the medical domain. Our implementation is built in the Spark distributed computing framework (Zaharia et al., 2010). For a discussion and comparison of universal schemas to other AKBC methods, we refer the reader to Section 2 of (Verga et al., 2015), which provides an excellent overview.

The contributions of this paper are as follows:

- a confirmation that the universal schema approach can be effective in a domain specific settings; and
- an exemplar of how straightforward AKBC methods can be applied using widely available compute platforms.

We begin with a description of our system followed by an set of initial experiments and then conclude.

2 System Description

Our CAT3-KB system consists of seven steps depicted in Figure 1: open information extraction, ontology ingestion, concept resolution, matrix construction, matrix factorization, matrix completion, and curation. We now describe these steps in turn.

Open IE ScienceDirect content consists of XML representations of articles. From this XML, we extract the plain text of articles. These articles are fed through a Spark-based reimplementation of ReVerb (Fader et al., 2011). At a high level, ReVerb identifies relation phrases by looking for text spans starting with a verb and ending with a preposition (e.g. "is the leading cause of"). Noun phrases before and after the span are used as the arguments of the subsequent relation instance (i.e. fact). We lemmatize all relations, and only keep relation types that have more than 5 distinct argument pairs and that occur more than 25 times. These parameters are adjustable. From the original 14 million articles (representing approximately 1 TB of text), 475 million facts are returned. This amount of data is an example of why it is helpful to employ a distributed computing framework such as Spark.

Ontology Ingestion To combine the lemmatized surface form relations from the prior step with an imported ontology, we first ingest the ontology into an annotation engine. We have developed the Solr Dictionary Annotator (SoDA)[1], a high performance dictionary based annotation engine for Spark. SoDA is specifically designed to support large dictionaries such as EMMeT. Because of the diversity of scientific content and the availability of large ontologies, performing concept recognition type tasks using a dictionary approach is often effective. Importantly, this approach allows us to adjust our knowledge base to different domains by ingesting different ontologies.

Concept Resolution In this stage, we run SoDA against the noun phrase arguments of the lemmatized surface form relation instances. It matches the arguments to concepts within the ingested ontology. This process includes fuzzy string matching against all the synonyms of the various concepts. The output of this step is a knowledge graph consisting of known concepts linked by both surface form relation instances as well as by relations from the ontology.

The concept resolution step reduces the 475 million Open IE facts to 46 million where both surface form arguments match EMMeT concepts. We add in the three million facts from EMMeT, giving us a medical knowledge base of 49 million facts.

Matrix Construction Following (Riedel et al., 2013), we construct a matrix from those 49 million facts. The rows are the pairs of arguments in each fact, and the columns are the lemmatized relations (or the known relations from EMMeT). The cells have binary values; one where the two arguments are linked by that relation, zero where not. As described later, our initial experiments use subsets of this matrix.

Matrix Factorization (Riedel et al., 2013) presents a number of models based on the universal schema representation. A key insight of that work was that these models could take advantage of techniques from collaborative filtering. We apply this insight directly and build a model using alternating

[1] https://databricks.com/blog/2016/02/10/how-elsevier-labs-implemented-dictionary-annotation-at-scale-with-apache-spark-on-databricks.html

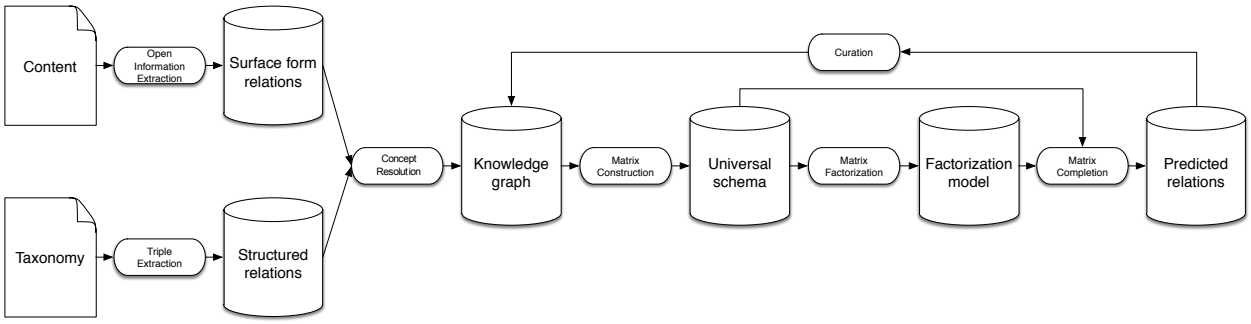

Figure 1: The Architecture of CAT3-KB

least squares (ALS) to approximate the given universal schema as the product of two lower rank matrices. Each relation is approximated by one factor and each concept-concept pair is approximated by another. In our setting, we use Spark's existing parallel implementation of an implicit feedback version of ALS with regularization. The number of iterations, regularization, and implicit feedback baseline confidence parameters are all set empirically. Those settings are given later in the experimental section.

From our initial observations, results appear to be robust to differing parameter settings but we have yet to perform a full sensitivity analysis. The output of this step are an Nxk and an Mxk factor matrices, where N is the number of concept-concept pairs and M is the number of lemmatized and EMMeT relations. The results described here are for k = 30.

Matrix Completion To generate a set of predicted relation instances, we approximate the original matrix by multiplying the two matrix factors. The resulting matrix is no longer binary; it has values between 0. and 1.0, inclusive. Therefore, we need to determine a threshold where a score generated by the model should be considered to constitute a relation instance.

To determine the threshold, we first perform tenfold cross validation on the input matrix. For each fold we factorize and complete the matrix, then report precision, recall and F1 measure between the original and completed matrix. This comparison is done for 11 different threshold values. Based on these results, we select the threshold that maximizes the F1 measure. After determining a threshold using the cross validation described above, we run matrix factorization across the entire input matrix and select those facts that score above the threshold.

An important point to make is that we are predicting new relations that do no explicitly appear in the literature or in the input ontology. To clarify, while just looking at the Open IE facts or input knowledge graph relations is of interest, our goal here is to predict new links not within that knowledge graph.

Curation The final step is to provide the set of predicted facts to experts for analysis and use in their application. These may be used to refine the input taxonomy through the addition of relations or inserted back into the input knowledge graph.

3 Initial Experiments

We have conducted an initial set of tests with small subsets of the input matrix so that we can easily examine the outputs manually and quickly iterate. Two subsets were created, based on concepts that appear frequently within in the Clinical Key search logs. The two concepts were "glaucoma" and "rheumatoid arthritis". For each concept, we create a small subset of the input graph by selecting the rows which have the concept in the arg1 position, and all columns that have any values in those rows. For glaucoma, this results in a matrix with 173 rows and 83 columns. The matrix is sparse[2] containing 356 existing relation instances. The columns for the EMMeT relations are relatively dense, the columns for the surface form relations are relatively sparse. We run 10 fold cross-validation on the input matrix and try 11 different threshold values in each fold. Results were averaged over the 10 folds and plotted in Figure 2. We achieve a maximum F1 measure of .71 at a threshold of 0.3, although F1 is not especially sensitive to the threshold. In our initial exper-

[2]The sparsity ratio is 1.

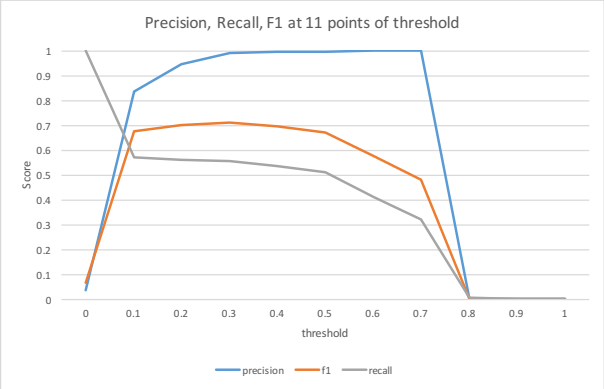

Figure 2: Performance in the glaucoma experiment

iments, we wanted to increase recall, so a threshold of 0.08 was chosen.

The complete glaucoma subset was factorized, completed, and thresholded. We found 22 relation instances that were not in the input knowledge graph. Figure 3 depicts several interesting examples.[3] We see synonymous surface forms (e.g. develop following) as well as relations between glaucoma and class concepts (e.g. age over 40). Note, that in Figure 3 we replace the concept ids with their preferred term from the EMMeT ontology.

For rheumatoid arthritis, the subset matrix had 258 rows by 70 columns and contained 465 known relation instances. The best performing threshold for F1 was slightly different than for the glaucoma example - a threshold of 0.2 gave the maximum F1 of 0.75. Here too we found a number of interesting and new relation instances. One that stood out in the result set was the ability to predict a new relation instance that used an ontology relation, namely (rheumatoid arthritis, emmet:isRiskFactor , amyloidosis). In the input set of relations, the link between arthritis and amyloidosis was that of causation and a potential complication.

4 Conclusion

This work presents CAT3-KB - a system for automated knowledge based construction based on universal schemas and implemented using the Spark distributed computing framework. A key attribute of our system is that it is primarily unsupervised. Using straightforward AKBC techniques (open information extraction, dictionary matching, and universal schemas), we were able to generate a 49 million fact knowledge graph which combined knowledge from medical text and from a large medical taxonomy. We were then able to perform link prediction with subsets of that graph allowing us to produce suggestions for ontology expansion.

We see the importance of this initial system, not in advancing new algorithms, but in showing the applicability and reusability of state-of-the-art AKBC methods over real world datasets.

There are a number of avenues for future work. From a use case perspective, we have presented this initial system to Elsevier's in-house ontology team. They were encouraged by the system and presented several use cases that we think are of interest to the AKBC community:

1. Querying over the ontology. This is a semantic search problem in which the aim is to find existing relations for a given query. Given an existing relation like "emmet:hasCause", the knowledge base can provide many surface forms that are synonymous. This will help ontology browsing and also adding synonyms to the ontology.

2. Provide evidence to confirm existing ontology relations. Medicine is particularly concerned with being based in evidence. Being able to link a particular fact with textual passages that support it is valuable for credibility.

3. Provide evidence for new relation instances not yet in an ontology. Being able to suggest that two concepts should be linked by one of the ontology's semantic types, and being able to back that up with paragraph level text passages, is a tremendous savings in the effort for ensuring the ontology is grounded in the literature.

From a system's perspective, our first step is to perform larger experiments representing the entire knowledge graph as a universal schema. We also intend to implement more sophisticated models based on the universal schema representation. Likewise, a

[3] It is important not to take the results too literally. Glaucoma is (probably) not the second leading cause of functional visual field loss, although it is the second leading cause of blindness in the US. This points out the need for human curation of the results before they are included in an ontology like EMMeT.

ARG1	REL	ARG2
glaucoma	developed many years after	chronic inflammation of uveal tract
glaucoma	develop following	chronic inflammation of uveal tract
glaucoma	can appear soon in	family history of glaucoma
glaucoma	can appear soon in	age over 40
glaucoma	is considered the second leading cause of	functional visual field loss
glaucoma	remains the second leading cause of	functional visual field loss

Figure 3: Examples of new relation instances for glaucoma

comparison to other distant suppression mechanisms would be informative. Additionally, we need to perform sensitivity analysis on our various parameter settings. We also would like to explore the applicability to other ontologies or combinations there of. In the longer term, a key question we have is ranking facts to be shown to experts.

In closing, we believe that AKBC methods have matured to the extent where they can be a key asset for ontologist in the maintenance and creation of large scale domain-specific ontologies.

Acknowledgements

We thank the reviewers for their important suggestions and guidance on directions for future work.

References

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics.

Carol Friedman, Pauline Kra, Hong Yu, Michael Krauthammer, and Andrey Rzhetsky. 2001. Genies: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(suppl 1):S74–S82.

Alexander G. Ororbia II, Jian Wu, and Lee C. Giles. 2014. Citeseerx: Intelligent information extraction and knowledge creation from web-based data. In *The 4th Workshop on AutomatedKnowledge Base Construction*, May.

Hoifung Poon, Chris Quirk, Charlie DeZiel, and David Heckerman. 2014. Literome: Pubmed-scale genomic knowledge base in the cloud. *Bioinformatics*.

Sebastian Riedel, Limin Yao, Benjamin M. Marlin, and Andrew McCallum. 2013. Relation extraction with matrix factorization and universal schemas. In *Joint Human Language Technology Conference/Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL '13)*, June.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 243–246, New York, NY, USA. ACM.

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2015. Multilingual relation extraction using compositional universal schema. *arXiv preprint arXiv:1511.06396*.

Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. 2010. Spark: Cluster computing with working sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, HotCloud'10, pages 10–10, Berkeley, CA, USA. USENIX Association.