End-to-End Learning for Answering Structured Queries Directly over Text

Paul Groth[1], Antony Scerri[2], Ron Daniel, Jr.[2], and Bradley P. Allen[2]

[1] University of Amsterdam p.groth@springer.com
[2] Elsevier Labs
{a.scerri,r.danel,b.allen}@elsevier.com

Abstract. Structured queries expressed in languages (such as SQL, SPARQL, or XQuery) offer a convenient and explicit way for users to express their information needs for a number of tasks. In this work, we present an approach to answer these directly over text data without storing results in a database. We specifically look at the case of knowledge bases where queries are over entities and the relations between them. Our approach combines distributed query answering (e.g. Triple Pattern Fragments) with models built for extractive question answering. Importantly, by applying distributed querying answering we are able to simplify the model learning problem. We train models for a large portion (572) of the relations within Wikidata and achieve an average 0.70 F1 measure across all models. We describe both a method to construct the necessary training data for this task from knowledge graphs as well as a prototype implementation.

1 Introduction

Database query languages (e.g. SQL, SPARQL, XQuery) offer a convenient and explicit way for users to express their information needs for a number of tasks including populating a dataframe for statistical analysis, selecting data for display on a website, defining an aggregation of two datasets, or generating reports.

However, much of the information that a user might wish to access using a structured query may not be available in a database and instead available only in an unstructured form (e.g. text documents). To overcome this gap, the area of *information extraction* (IE) specifically investigates the creation of structured data from unstructured content [15]. Typically, IE systems are organized as pipelines taking in documents and generating various forms of structured data from it. This includes the extraction of relations, the recognition of entities, and even the complete construction of databases. The goal then of IE is not to answer queries directly but first to generate a database that queries can be subsequently executed over.

In the mid-2000s, with the rise of large scale web text, the notion of combining information extraction techniques with relational database management systems emerged [4, 10] resulting in what are termed *text databases*. Systems like Deep Dive [22] InstaRead [9], or Indrex [11], use database optimizations within tasks such as query planning to help decide when to perform extractions. While, in some cases, extraction of data can be performed at runtime, data is still extracted to an intermediate database before the query is answered. Thus, all these approaches still require the existence of a structured database to answer the query.

In this paper, we present an approach that **eliminates the need to have an intermediate database in order to answer structured database queries over text**. This is essentially the same as treating the text itself as the store of structured data. Using text as the database has a number of potential benefits, including being able to run structured queries over new text without the need for a-priori extraction; removing the need to maintain two stores for the same information (i.e. a database and a search index); eliminating synchronization issues; and reducing the need for up-front schema modeling. [3] provides additional rationale for not pre-indexing "raw data", although they focus on structured data in the form of CSV files.

Our approach builds upon three foundations: 1. the existence of large scale publicly available knowledge bases (Wikidata) derived from text data (Wikipedia); 2. recent advances in end-to-end learning for extractive question answering (e.g. [21]); 3. the availability of layered query processing engines designed for distributed data (e.g. SPARQL query processes that work over Triple Pattern Fragment [25] servers).

A high-level summary of our approach is as follows. We use a publicly-available knowledge base to construct a parallel corpus consisting of tuples each which is made up of a structured slot filling query, the expected answer drawn from the knowledge base, and a corresponding text document in which we know the answer is contained. Using this corpus, we train neural models that learn to answer the given structured query given a text document. This is done on a per relation basis. These models are trained end-to-end with no specific tuning for each query. These models are integrated into a system that answers queries expressed in a graph query language directly over text with no relational or graph database intermediary.

The contributions of this paper are:

- an approach, including training data generation, for the task of answering structured queries over text;
- models that can answer slot filling queries for over 570 relations with no relation or type specific tuning. These models obtain on average a 0.70 F1 measure for query answering.
- a prototype system that answers structured queries using triple pattern fragments over a large corpus of text (Wikipedia).

The rest of this paper is organized as follows. We begin with an overview of the approach. This is followed by a description of the training data. Subsequently, we describe the model training and discuss the experimental results. After which, we present our prototype system. We end the paper with a discussion of related and future work.

2 Overview

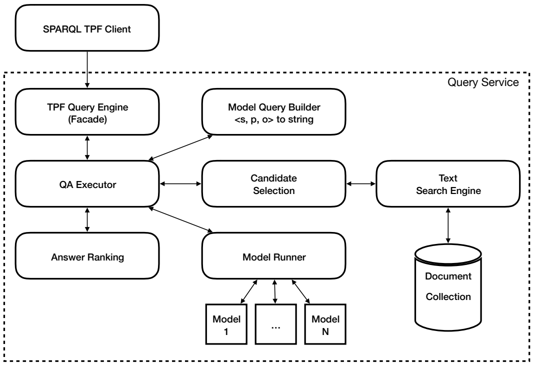


Fig. 1: Components of the overall system for structured query answering over text.

Our overall approach consists of several components as illustrated in Figure 1. First, structured queries as expressed by SPARQL [7] are executed using a Triple Pattern Client (SPARQL TPF Client). Such a client breaks down a more complex SPARQL query into a series of triple patterns that are then issued to a service. Triple patterns are queries of the form subject, predicate, object, where each portion can be bound to an identifier (i.e. URI) or a variable.[3] Within the service, the execution engine (QA Executor) first lexicalizes the given identifiers into strings using the Model Query Builder component. For example, this component would translate an identifier like https://www.wikidata.org/wiki/Q727 into the string form "Amsterdam". These queries are then issued to a candidate selection component. This component queries a standard text search engine to find potential documents that could contain the answer to the specified query.

The candidate documents along with the lexicalized queries are provided to a model runner which issues these to models trained specifically to bind the variable that is missing. That is given a query of the form $< s, p, ?o >$ where s and o are bound and o is the variable, there would be specific models trained to extract $?o$ from the provided document. For example, given the query (:Amsterdam :capital_of $?o$) we would have models that know how to answer queries of where the type of the subject is City and the property is capital_of. Likewise, there would be models of that are able to answer queries of the form $<?s, p, o >$ and so on. Each model is then asked to generate a binding of the variable. Note that the bindings generated by the models are strings. The results of each model are then ranked (Answer Ranking). Using a cut-off, the results are then translated back into identifier space and returned to the client.

[3] Objects can also be bound to a literal.

A key insight of our approach is that by breaking down complex queries into triple patterns we can simplify the queries that need to be answered by the learned models.

Our approach relies on the construction of models that are able to extract potential candidate answers from text. Following from [5] and [13], we cast the problem in terms of a question answering task, where the input is a question (e.g. entity type + relation) and a document and the output is answer span within the document that binds the output. To learn these sorts of models we construct training data from knowledge graphs that have a corresponding representation in text. In the next section, we go into detail about the construction of the necessary training data.

3 Training Data Construction

Our training data is based on the combination of Wikidata and Wikipedia. Wikidata is a publicly accessible and maintained knowledge base of encyclopedic information [27]. It is a graph structured knowledge base (i.e. a knowledge graph) describing entities and the relations between them. Every entity has a globally unique identifier. Entities may also have attributes which have specific datatypes. Entities have may have more than one type. Relations between entities may hold between differing entity types.

Wikidata has a number of properties that make it useful for building a corpus to learn how to answer structured queries over text. First, and perhaps most importantly, entities have a parallel description with Wikipedia. By our count, Wikidata references 7.7 million articles in the English language Wikipedia. Thus, we have body of text which will also most likely contain answers that we retrieve from Wikidata. Second, every entity and relation in Wikidata has a human readable label in multiple languages. This enables us to build a direct connection between the database and text. Third, Wikidata is large enough to provide for adequate training data in order to build models. Finally, Wikidata provides access to their data in a number of ways including as a SPARQL endpoint, a triple patterns fragment endpoint and as a bulk RDF download. While we use Wikidata, we believe that our approach can be extended to any knowledge graph that has textual sources.

Using this input data we generate datasets of the form: [QUERY; ANSWER; TEXT IN WHICH THE QUERY IS ANSWERED]. As previously mentioned, complex queries can be expressed as a series of graph patterns. Thus, the queries we consider are graph patterns in which two of the variables are bound (e.g. :NEW_ENGLAND_PATRIOTS :PLAY ?x). We term these *slot filling* queries as the aim is to bind one slot in the relation (i.e. the subject or the object). While we do not test graph patterns where the predicate is the variable, the same approach is also applicable. In some sense, one can think of this as generating data that can be used to build models that act as substitute indexes of a database.

Our construction method loops through all of the predicates (i.e. relations) in the dataset. It determines the frequency with which a predicate connects different types of entities. This is essential as large knowledge graphs can connect many different types using the same predicate. Thus, examples from different types of subjects and objects are needed to capture the semantics of that predicate. Using the most frequently occurring pairs of entity types for a predicate, the algorithm then retrieves as many example triples as possible where the subject and object of the triple are instances of the connected types - up to a given maximum threshold. Thresholding is used to help control the size of the training data.

Each triple is then used to generate a row of training data for learning how to answer graph pattern queries that contain the given predicate. To connect the graph pattern queries, which are expressed using entity IRIs to the plain text over which it should be answered, each of the components of the triple is lexicalized. The lexicalized subject and predicate of each triple are concatenated together to form a textual query and use the lexicalized object as the answer. (Note, this is trivally modified for the (?s, p, o case). We then retrieve the text describing the subject. We assume that the text contains some reference to the object under consideration.

The location of that reference which we term an anchor is computed by the given anchor function. For simplicity, in our implementation, we locate the first instance of the answer in the text. This may not always represent an instance of the answer's lexical form which is located in an expression which answers the specific question form. More complex implementations could use different heuristics or could return all possible anchor locations.

We apply the algorithm to the combination of Wikipedia and Wikidata dumps[4]. We attempted to obtain training data for all 1150 predicates in Wikidata that associate two entities together. At this time, we do not treat predicates that connect entities to literals. This is left for future work. We limited the extraction to the top 20 entity type pairs per predicate, and limited each type pair to 300 examples). Thus, there is a maximum yield of 6000 examples per predicate. We then apply the following cleaning/validation to the retrieved examples. First, we drop examples where there is no Wikipedia page. Second, we ensure that the answer is present in the Wikipedia page text. Finally, in order to ensure adequate training data we filter out all models with less than 30 examples. Note that this means that we have differing amounts of training data per predicate. After cleaning, we are able to obtain training data for 572 predicate for the setting in which the object is the variable/answer. We term this the SP setting. On average we have 929 examples per predicate with a maximum number of examples of 5477 and a minimum of 30 examples. The median number of examples is 312. In the setting in which the subject is the variable / answer we are trying to extract, enough data for 717 predicates is obtained. This is because the subject answer is more likely to appear in the Wikipedia page text. We term this the PO setting.

[4] Specifically we used Wikipedia 2018-08-20 (enwiki-20180820-pages-articles-multistream.xml.bz2) and Wikidata 2018-08-29.

4 Models

Based on the above training data, we individual train models for all predicates using the Jack the Reader framework [5]. We use two state-of-the-art deep learning architectures for extractive question answering, namely, FastQA [28] and the implementation provided by the framework, JackQA. Both architectures are interesting in that while they perform well on reading comprehension tasks (e.g. SQuAD [20]) both architectures try to eliminate complex additional layers and thus have the potential for being modified in the future to suit this task. Instead of describing the architectures in detail here, we refer the reader to corresponding papers cited above. We also note that the Jack the Reader configuration files provide succinct descriptions of the architectures, which are useful for understanding their construction.

To improve performance both in terms of reducing training time and to reduce the amount of additional text the model training has to cope with, we applied a windowing scheme. This is because longer text is normally associated with greater issues when dealing with sequence models. Our scheme takes a portion of the text around the answer location chosen from the Wikipedia content. We now describe the following parameters for each architecture.

FastQA All text is embedded using pre-trained GloVe word embeddings [17] (6 billion tokens, and 50 dimensions). We train for 10 epochs using a batch size of 20. We constrain answers to be a maximum of 10 tokens and use a window size of 1000 characters. The answer layer is configured to be bilinear. We use the ADAM optimizer with a learning rate of 0.11 and decay of 1.0.

JackQA Here we embed the text using pre-trained GloVe word embeddings (840 billion tokens and 300 dimensions). We use the default JackQA settings. We use a window size of 3000 characters. The batch sizes were 128/96/64 for three iterative runs. The subsequent runs with smaller batch sizes were only run if the prior iteration failed. We specified a maximum number of 20 epochs.

Baseline In addition to the models based on neural networks, we also implemented a baseline. The baseline consisted of finding the closest noun phrase to the property within the Wikipedia page and checking whether the answer is contained within that noun phrase.

Note, we attempted to find functional settings that worked within our available computational constraints. For example, FastQA requires more resources than JackQA in relation to batch size , thus, we chose to use smaller embeddings and window size in order to maintain a “good” batch size.

We use 2/3 of the training data for model building and 1/3 for testing. Data is divided randomly. Training was performed using an Amazon EC2 p2.xlarge[5] box. It took 23 hours for training of FastQA models, which included all models for all predicates even when there were too few training samples. For JackQA, the window was increased to 3000 characters, and multiple training sessions were required, reducing the batch size each time to complete the models which not

[5] 1 virtual GPU - NVIDIA K80, 4 virtual CPUs, 61 GiB RAM

finish from earlier runs, in all three passes were required with 128, 96 and 64 batch size respectively. Total training time was 81 hours.

Note that we train models for the setting where the subject and predicate are bound but the object is not. We also use the FastQA architecture to build models for the setting where the subject is treated as the variable to be bound.

5 Experimental Results and Analysis

Table 1 one reports the average F1 measure across all models as well as the baseline. This measure takes into account the overlap of the identified set of tokens with the gold standard answer controlling for the length of the extracted token. By definition, the baseline only generates such overlap scores.

Model	Model Count	mean	std	min	max	25%	50%	75%
JackQA - SP	572	0.70	0.24	0.0	1.0	0.54	0.77	0.89
FastQA - SP	572	0.62	0.24	0.0	1.0	0.43	0.65	0.80
FastQA - PO	717	0.89	0.10	0.4	1.0	0.85	0.92	0.96
Baseline	407	0.15	0.17	0.0	0.86	0.03	0.08	0.20

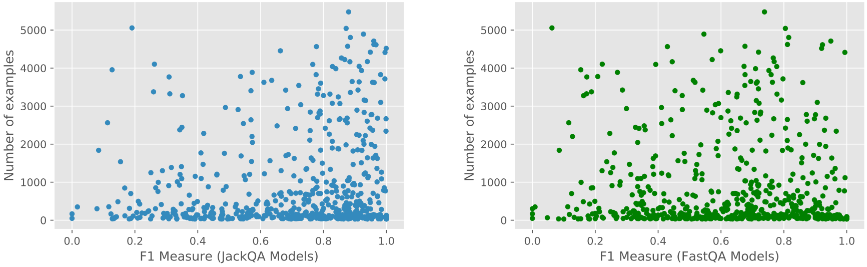
Table 1: F1 results across all models and the baseline

Table 2 reports the average exact match score over all models. This score measures whether the model extracts the exact same string as in the gold standard. For reference, both tables also reports the total number of models trained (Model Count), which is equivalent to the training data provided. The Model Count for the baseline is equivalent to the number of predicates for which the baseline method could find an answer for.

Model	Model Count	mean	std	min	max	25%	50%	75%
JackQA - SP	572	0.64	0.26	0.0	1.0	0.44	0.71	0.86
FastQA - SP	572	0.55	0.25	0.0	1.0	0.36	0.57	0.74
FastQA - PO	717	0.83	0.14	0.1	1.0	0.75	0.86	0.94

Table 2: Exact results

Figure 3 plots individual model performance against the size of the training data given. Overall, models based on deep learning notably outperform the baseline models on average. Additionally, using these deep learning based approaches we are able to create models that answer queries for 160 additional properties over the baseline. In terms of analysis, first, we wanted to see if there was a correlation between the amount of training data and the performance of a model. Using the data presented in Figure 3, we fit a linear regression to it. We found no statistically significant correlation ($R^2 = 0.37$). The model architectures show



Models trained using Jack QA

Models trained using Fast QA.

Fig. 3: Plot of individual model performance vs. training data size. All 572 models are shown for the SP setting.

strong correlation in performance. The R^2 value being 0.97 in the case of the F1 measure and 0.96 for the Exact measure. This suggests that the performance is primarily a factor of the underlying kind of data. More details are provided in Appendix A

6 Prototype

To understand whether this approach is feasible in practice, we implemented a prototype of the system outlined in Figure 1. For the triple pattern fragment facade we modify Piccolo, an open source triple pattern fragments server to replace its in-memory based system with functions for calling out to our QA answering component. The facade also implements a simple lexicalization routine. The query answering component is implemented as a Python service and calls out to an Elasticsearch search index where documents are stored. The query answering component also pre-loads the models and runs each model across candidate documents retrieved by querying elastic search. We also specify a max number of candidate documents to run the models over. Currently, we execute each model sequentially over all candidate documents. We then chose the top set of ranked answers given the score produced by the model. Note that we can return multiple bindings for the same ranked results. We made some preliminary timing estimates of a query. It takes on the order of 10 seconds to provide results for a single triple pattern query. This is surprisingly good given the fact that we execute models sequentially instead of in parallel. Furthermore, we execute the models over the entirety of the Wikipedia article. Our own anecdotal experience shows that question answering models are both faster and produce more accurate results when supplied with smaller amounts of text. Thus, there is significant room for optimizing query performance with some simple approaches including parallelizing models, chunking text into smaller blocks, and limiting the number of models executed to those that are specific for the triple pattern. Furthermore, it is straightforward to issue triple pattern fragment query requests over multiple running instances [25]. One could also implement more complex sharding mechanisms designed for triple tables [1]. Overall, the prototype gives us confidence that this sort of system could be implemented practically.[6]

7 Related Work

Our work builds upon and connects to a number of existing bodies of literature. The work on information extraction is closely related. [15] provides a recent survey of the literature in this area specifically targeted to the the problems of extracting and linking of entities, concepts and relations. One can view the models that we build as similar to distantly supervised relation extraction approaches [16, 23], where two mentions of entities are found in text and the context around those mentions is used to learn evidence for that relation. Recent approaches have extended the notion of context [18] and applied neural networks to extract relations [30, 6].

The closest work to ours in the information extraction space is [14] where they apply machine comprehension techniques to extract relations. Specifically, they translate relations into templated questions - a process they term querification. For example, for the relation spouse(x,y) they created a series of corresponding question templates such as “Who is x married to?”. These templates are constructed using crowdsourcing, where the workers are provided a relation, example sentence and asked to produce a question template. This dataset is used to train a BiDAF-based model [21] and similar to our approach they address slot filling queries where the aim is to populate one side of the relation. While we apply a similar technique, our approach differs in a two key aspects. First, we target a different task, namely, answering structured queries. Second, we do not generate questions through question templates but instead build the questions out of the knowledge base itself.

Like much of the work in this space our approach is based on a large scale parallel corpus. Of particular relevance to our task are the WikiSQL and WikiReading corpora. WikiSQL [31] provides a parallel corpus that binds SQL queries to a natural language representation. The task the dataset is used for is to answer natural language questions over SQL unlike ours which is to answer SQL-like queries over text. SQLWikiReading [8] like our approach extracts a corpus from Wikidata and Wikipedia in order to predict the value of particular properties. Another corpus of note is ComplexWebQuestions [24], which pairs complex SPARQL queries with natural language queries. Importantly, it looks at the compositionality of queries from smaller units. Like WikiSQL, it looks at answering natural language queries over databases. In general, we think our approach in also specifying an extraction procedure is a helpful addition for applying corpus construction in different domains.

As mentioned in the introduction, text databases, where information extraction is combined with databases are also relevant. Our system architecture was

[6] We also integrated the prototype with Slack.

inspired by the pioneering work of [10]. In that work, a search index is used to first locate potential documents and then information extraction techniques are applied to the selected documents to populate a database. Our approach differs in two key aspects. First, instead of populating a database our system substitutes the indexes of the database with models. Second, we use distributed query techniques in order to process complex queries on the client side. Recent work [12] uses deep learning based approaches to perform information extraction during database query execution specifically for entity disambiguation. Similar to other work in this area, and unlike ours, they integrate the information extraction within the database engine itself.

Finally, there is a long history of mixing information retrieval and database style queries together. For example, for the purposes of querying over semistructured data [2]. [19] provides an accessible introduction to that history. While our system is designed to answer database queries one can imagine easily extending to the semistructured setting.

8 Conclusion & Future Work

In this work, we have explored the notion of answering database queries over text absent the need for a traditional database intermediary. We have shown that this approach is feasible in practice by combining machine comprehension based models with distributed query techniques.

There are a number of avenues for future work. In the short term, the developed models could be expanded to include extracting properties as well as subjects and objects. We also think that joint models for all triple pattern predictions is worth exploring. One would also want to extend the supported queries to consider not only relationships between entities but also to the attributes of entities. Our current lexicalization approach is also quite simple and could be improved by considering it as the inverse of the entity linking problem and applying those techniques or applying summarization approaches [26]. In this work, we used model architectures that are designed for answering verbalized questions and not database queries. Modifying these architectures may also be a direction to obtain even better performance. Obviously more extensive experimental evaluations would be of interest, in particular, extending the approach to other knowledge bases and looking more deeply at query result quality.

In the long term, the ability to query over all types of data whether images, structured data or text has proven useful for knowledge bases [29]. Extending our concept to deal with these other datatypes could be powerful -making it easy to perform structured queries over unstructured data while minimizing information extraction overhead. In general, we believe that structured queries will continue to be a useful mechanism for data professionals to both work with data and integrate information into existing data pipelines. Hence, focusing on automated knowledge base construction from the query vantage point is an important perspective.

References

1. Abdelaziz, I., Harbi, R., Khayyat, Z., Kalnis, P.: A survey and experimental comparison of distributed sparql engines for very large rdf data. Proceedings of the VLDB Endowment **10**(13), 2049–2060 (2017)
2. Abiteboul, S.: Querying semi-structured data. In: International Conference on Database Theory. pp. 1–18. Springer (1997)
3. Alagiannis, I., Borovica, R., Branco, M., Idreos, S., Ailamaki, A.: Nodb: efficient query execution on raw data files. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. pp. 241–252. ACM (2012)
4. Cafarella, M.J., Re, C., Suciu, D., Etzioni, O., Banko, M.: Structured querying of web text. In: 3rd Biennial Conference on Innovative Data Systems Research (CIDR), Asilomar, California, USA (2007)
5. Dirk Weissenborn, Pasquale Minervini, T.D.I.A.J.W.T.R.M.B.J.M.T.D.P.S.S.R.: Jack the Reader A Machine Reading Framework. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL) System Demonstrations (July 2018), https://arxiv.org/abs/1806.08727
6. Glass, M., Gliozzo, A., Hassanzadeh, O., Mihindukulasooriya, N., Rossiello, G.: Inducing implicit relations from text using distantly supervised deep nets. In: International Semantic Web Conference. pp. 38–55. Springer (2018)
7. Harris, S., Seaborne, A., Prudhommeaux, E.: Sparql 1.1 query language. W3C recommendation **21**(10) (2013)
8. Hewlett, D., Lacoste, A., Jones, L., Polosukhin, I., Fandrianto, A., Han, J., Kelcey, M., Berthelot, D.: WIKIREADING: A novel large-scale language understanding task over Wikipedia. In: Proceedings of the The 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016) (2016)
9. Hoffmann, R., Zettlemoyer, L., Weld, D.S.: Extreme extraction: Only one hour per relation. arXiv preprint arXiv:1506.06418 (2015)
10. Jain, A., Doan, A., Gravano, L.: Sql queries over unstructured text databases. In: Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. pp. 1255–1257. IEEE (2007)
11. Kilias, T., Löser, A., Andritsos, P.: Indrex: In-database relation extraction. Information Systems **53**, 124–144 (2015)
12. Kilias, T., Löser, A., Gers, F.A., Koopmanschap, R., Zhang, Y., Kersten, M.: Idel: In-database entity linking with neural embeddings. arXiv preprint arXiv:1803.04884 (2018)
13. Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., Socher, R.: Ask me anything: Dynamic memory networks for natural language processing. In: International Conference on Machine Learning. pp. 1378–1387 (2016)
14. Levy, O., Seo, M., Choi, E., Zettlemoyer, L.: Zero-shot relation extraction via reading comprehension. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). pp. 333–342 (2017)
15. Martinez-Rodriguez, J., Hogan, A., Lopez-Arevalo, I.: Information extraction meets the semantic web: A survey. Semantic Web Journal (2018)
16. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: 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. pp. 1003–1011. Association for Computational Linguistics (2009)

17. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014), http://www.aclweb.org/anthology/D14-1162
18. Quirk, C., Poon, H.: Distant supervision for relation extraction beyond the sentence boundary. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. vol. 1, pp. 1171–1182 (2017)
19. Raghavan, S., Garcia-Molina, H.: Integrating diverse information management systems: A brief survey. Bulletin of the Technical Committee on Data Engineering p. 44 (2001)
20. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2383–2392 (2016)
21. Seo, M., Kembhavi, A., Farhadi, A., Hajishirzi, H.: Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603 (2016)
22. Shin, J., Wu, S., Wang, F., De Sa, C., Zhang, C., Ré, C.: Incremental knowledge base construction using deepdive. Proceedings of the VLDB Endowment **8**(11), 1310–1321 (2015)
23. Surdeanu, M., Tibshirani, J., Nallapati, R., Manning, C.D.: Multi-instance multi-label learning for relation extraction. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. pp. 455–465. Association for Computational Linguistics (2012)
24. Talmor, A., Berant, J.: The web as a knowledge-base for answering complex questions. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). vol. 1, pp. 641–651 (2018)
25. Verborgh, R., Sande, M.V., Hartig, O., Herwegen, J.V., Vocht, L.D., Meester, B.D., Haesendonck, G., Colpaert, P.: Triple pattern fragments: A low-cost knowledge graph interface for the web. Web Semantics: Science, Services and Agents on the World Wide Web **37-38**, 184 – 206 (2016). https://doi.org/https://doi.org/10.1016/j.websem.2016.03.003, http://www.sciencedirect.com/science/article/pii/S1570826816000214
26. Vougiouklis, P., Elsahar, H., Kaffee, L.A., Gravier, C., Laforest, F., Hare, J., Simperl, E.: Neural wikipedian: Generating textual summaries from knowledge base triples. Journal of Web Semantics (2018). https://doi.org/https://doi.org/10.1016/j.websem.2018.07.002, http://www.sciencedirect.com/science/article/pii/S1570826818300313
27. Vrandečić, D.: Wikidata: A new platform for collaborative data collection. In: Proceedings of the 21st International Conference on World Wide Web. pp. 1063–1064. ACM (2012)
28. Weissenborn, D., Wiese, G., Seiffe, L.: Making neural qa as simple as possible but not simpler. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). pp. 271–280 (2017)
29. Wu, S., Hsiao, L., Cheng, X., Hancock, B., Rekatsinas, T., Levis, P., Ré, C.: Fonduer: Knowledge base construction from richly formatted data. In: Proceedings of the 2018 International Conference on Management of Data. pp. 1301–1316. ACM (2018)
30. Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J.: Relation classification via convolutional deep neural network. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 2335–2344 (2014)

31. Zhong, V., Xiong, C., Socher, R.: Seq2sql: Generating structured queries from natural language using reinforcement learning. CoRR **abs/1709.00103** (2017)

A Individual Model and Error Analysis

We looked more deeply at performance for individual models for a given property. Table 3 shows the highest performing models. We find some consistent patterns. First, properties that have specific value constraints within Wikidata generate good results. For example, the "crystal system" property needs to have one of 10 values (e.g cubic crystal system, quasicrystal, amorphous solid). Likewise, the "coolant" property needs to be assigned one of fourteen different values (e.g. water, oil, air). This is also true of "discovery method", which oddly enough is actually defined as the the method by which an exoplanet is discovered. This is also a feature of properties whose values come from classification systems (e.g. "Kppen climate classification" and "military casualty classification").

A second feature that seems to generate high performing models are those that refer to common simple words. For example, the "source of energy" property takes values such as "wind" or "human energy".

Lastly, simple syntactic patterns seem to be learned well. For example, the property "birthday", which links to entities describing a month, day combination (e.g. November 8) which is thus restricted to a something that looks like a month string followed by one or two numerical characters. Likewise, the expected value for the property "flag" often appears directly in text itself. That is the correct answer for the query "Japan flag" is "flag of Japan", which will appear directly in text.

Property	Fast QA	Fast QA	Jack QA	Jack QA	Training
	F1	Exact	F1	Exact	Data Size
birthday	0.95	0.91	1.0	1.0	32
flag	0.98	0.88	1.0	1.0	50
league points system	1.00	1.00	1.0	1.0	90
discovery method	0.98	0.91	1.0	1.0	69
source of energy	0.94	0.94	1.0	1.0	50
military casualty classification	1.00	1.00	1.0	1.0	92
topic's main category	0.99	0.91	1.0	1.0	31
Kppen climate classification	1.00	1.00	1.0	1.0	34
coolant	0.98	0.98	1.0	1.0	128
crystal system	0.96	0.87	1.0	1.0	43

Table 3: Highest 10 performing models in the SP setting as determined by F1 measures from models trained using the Jack QA architecture.

We also look at the lowest performing models, shown in Table 4 to see what is difficult to learn. Ratings for films (e.g. Australian Classification, RTC film rating, EIRIN film rating) seem extremely difficult to learn. Each of these properties expect values of two or three letters (e.g. PG, R15+, M). The property "blood type" also has the same form. It seem that using character level embeddings may worked better in these cases.

The property "contains administrative territorial entity " is an interesting case as there are numerous examples. This property is used within Wikidata to express the containment relation in geography. For example, that county contains a village or a country contains a city. We conjecture that this might be difficult to learn because the sheer variety of linkages that this can express making it difficult to find consistencies in the space. A similar issue could be present for properties such as "voice actor" and "cast member" where the values can be essentially any person entity. Similarly, "polymer of" and "species kept" both can take values that come from very large sets (e.g. all chemical compounds and all species). It might be useful for the model to be provided specific hints about types (i.e. actors, chemicals, locations) that may allow it to find indicative features.

Property	Fast QA	Fast QA	Jack QA	Jack QA	Training
	F1	Exact	F1	Exact	Data Size
Australian Classification	0.00	0.00	0.00	0.00	48
RTC film rating	0.00	0.00	0.00	0.00	167
EIRIN film rating	0.01	0.01	0.02	0.02	349
blood type	0.00	0.00	0.08	0.08	302
contains administrative territorial entity	0.09	0.06	0.08	0.07	1838
voice actor	0.12	0.11	0.11	0.09	2562
species kept	0.11	0.03	0.12	0.03	354
best sprinter classification	0.19	0.18	0.12	0.11	165
cast member	0.15	0.14	0.13	0.11	3955
polymer of	0.18	0.08	0.13	0.08	38

Table 4: Lowest 10 performing models in the SP setting as determined by F1 measures from models trained using the Jack QA architecture.