Cleaning Noisy Knowledge Graphs

Ankur Padia

Computer Science and Electrical Engineering
University of Maryland, Baltimore County
Baltimore. MD 21250 USA
ankurpadia@umbc.edu

Abstract. My dissertation research is developing an approach to identify and explain errors in a knowledge graph constructed by extracting entities and relations from text. Information extraction systems can automatically construct knowledge graphs from a large collection of documents, which might be drawn from news articles, Web pages, social media posts or discussion forums. The language understanding task is challenging and current extraction systems introduce many kinds of errors. Previous work on improving the quality of knowledge graphs uses additional evidence from background knowledge bases or Web searches. Such approaches are difficult to apply when emerging entities are present and/or only one knowledge graph is available. In order to address the problem I am using multiple complementary techniques including entity-linking, common sense reasoning, and linguistic analysis.

1 Problem Statement

Problem Statement: Given a knowledge graph of entities and relations extracted from text along with optional source documents, provenance information, and confidence scores, how can we narrow down, identify and explain likely errors.

Many information extraction (IE) systems have been developed to extract information from multiple sources like spreadsheets, Wikipedia Infoboxes, discussion forum, and news articles [5, 4, 17, 10]. Fact are extracted following an IE paradigm, either Open IE or model IE. Model IE based system populate an ontology while Open IE extract possible facts without a target ontology. Extracted ontology is serialized in a standard representation like RDF. However, the task to extract information is challenging and current IE systems make many mistakes. Given the importance of knowledge graphs (KG) for downstream applications like Named Entity Recognizer/Typing, Semantic Role Labeling noise present in the KG is sipped in during distant supervision hurting the performance of the system.

Errors are caused by many factors, including ambiguous, conflicting, erroneous and redundant information [17], or can be due to reasons like schema violations, presence of outliers, and misuse of datatype properties [12, 18]. If the validity of an atomic fact is suspect, we call the fact *dubious*. Aim of my thesis is

to build a pipeline to identify and explain such dubious facts present in a knowledge graph. In this thesis, I plan to consider large general purpose knowledge graphs which are generated an IE system.

One of the solution is to use a confidence score to filter out correct candidates facts. However, such a technique can assign high confidence to incorrect fact and vice versa. For example, consider the fact from Never Ending Language Learner (NELL) [11] with great confidence *aaron-mckie is an actor*[1]; in reality he is a coach[2].

In this dissertation, I plan to exploring following aspects of the problem: (1) How can a subset of dubious facts be effectively narrowed down; (2) How accurately can we confirm that a candidate is, in fact, incorrect; and (3) explain why the fact is incorrect. I plan to address and answer these questions with several complementary approaches, including linguistic analysis, common-sense reasoning, and entity linking.

2 Relevancy

Considering the importance of KG in real world applications, in this section we perform quality assessment of two IE system: (1) Never Ending Language Learner (NELL)[11], and (2) Kelvin[10]. An automatic evaluation of the system can be a research problem in itself, we evaluate the quality using confidence score and expert engineered queries, respectively. Aim of the assessment is to highlight that irrespective of the corpus size, targeted domains, and training methodology, systems make mistakes jeopardizing its utility in downstream applications.

Web scale KG: **NELL** [11] is a semi-supervised, ontology driven, iterative system that extracts facts from a Web corpus of more than one billion documents. In each iteration, NELL learns new facts and assigns confidence score to it using previous facts and available evidence. For example, as of the 990^{th} iteration, there were approximately 97 million facts. with ten million had confidence of 0.8 and above and nearly 77 million facts were in low range of 0.5 and 0.6. Confidence scores are used to identify correct facts. However, for an IE system its possible to have high confidence for incorrect fact and visa versa. Based solely on confidence score, it's evident that NELL contains many dubious facts. Evaluating the correctness of such a large knowledge graph is a research problem. To better evaluate learning capability of an IE system, we additionally considered a more controlled system – Kelvin.

ColdStart based KG: **Kelvin** [10] is an unsupervised information extraction system that extracts facts from text to populate an ontology. It was initially developed to take part in the NIST TAC ColdStart Knowledge Base Population (CS-KBP) task[3]. The intention of TAC Coldstart KBP was to encourage build-

[1] http://rtw.ml.cmu.edu/rtw/kbbrowser/actor:aaron_mckie

[2] https://en.wikipedia.org/wiki/Aaron_McKie. A Google search query of 'Aaron-McKie actor' mentions his name in a number of IMDB pages since he was the subject of a documentary film, which may have led to his being classified as an actor

[3] http://tac.nist.gov/

ing knowledge base from scratch using set of text document without accessing external resources like Web search engines or Wikipedia. The choice to analyze Kelvin, beside other seven participants (Stanford, UMass, and others), is motivated by easy access to internal HLTCOE[4] resources and its relatively better extraction performance. For TAC 2015, Kelvin learned 3.2 million facts from total of 50K documents related to local news articles and web documents. Queries with given subject and relation and missing object were engineered by expert from Linguistic Data Consortium (LDC) were used to evaluate the accuracy of information extraction. On an average, due to mistake in understanding of relation among entities, Kelvin scored less then 30 F1 points making Kelvin a 2^{nd} ranked system with a small difference from first rank.

Reasons for errors: An extracted fact can be incorrect due to multiple reasons in addition to those discussed in [17]. Additional factors include the following: (1) the choice of natural language processing libraries; (2) design consideration of internal system components like classifiers and cross-document co-reference resolution systems; (3) bad entity type assignments; (4) poor in-document co-reference resolution; (5) missing mentions or extracting incorrect mentions from the text; (6) learning techniques, semi-supervised (as in NELL) or unsupervised (as in Kelvin); (7) heuristics employed; (8) choice of IE paradigm, model based or open; (9) quality of inferencing either using rules or statistical techniques; and (10) the nature of the text used to extract information.

With the success of the proposed approach, I plan to contribute algorithms to improve the quality of the KG increasing its utility in downstream applications.

3 Research Question(s)

Narrow Can potential incorrect fact candidates be effectively identified
Confirm How well can we confirm that a candidate is, in fact, incorrect
Explain How to rank or identify reasons why a fact might be wrong

In order to manage incorrect facts in a knowledge graph I am using a three stage pipeline: *Narrow Down*, *Confirm*, and *Explain*. The *Narrow Down* is a triage stage that selects a subset of all facts that are likely to be wrong, an important step for very large KGs. *Confirm* is more granular then Narrow Down and aims to classify the suspected facts as correct or incorrect. For each of the latter, *Explain* provides a human-understandable explanation including the sources from which the information was extracted and the likely cause of the error.

4 Hypotheses

H1: Narrow Down: Metric like confidence or frequency count are insufficient to identify likely possible facts
H2: Confirm: Better classification can be made when information at multiple levels, i.e type and instance level, is taken into consideration (as in Sec. 6)

[4] http://hltcoe.jhu.edu/

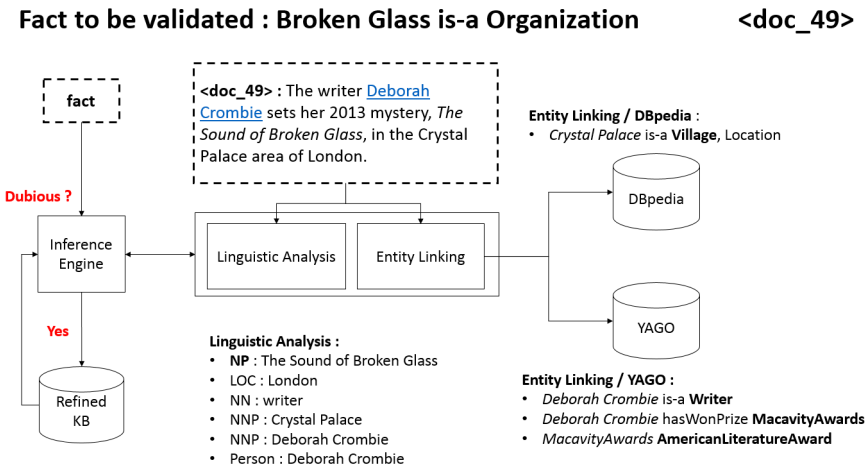

Fig. 1: Framework and a real world example to demonstrate benefits of entity linking with linguistic analysis of the text to help maintain incorrect facts.

H3: Explain: Structure-based techniques are better candidates to locate provenance information in web pages

Narrow Down is an important phase in cleaning knowledge bases. By default, all the facts could be considered dubious, but processing large KGs with millions of facts will be demanding and redundant. Techniques like frequency count of a predicate in the text and/or other KBs is of limited help, as it does not capture interactions between different predicates e.g co-occurrence or three-way interaction.

Confirm helps to determine the credibility of the given facts. Previous approaches like [17] have addressed this question to some extent with assumptions that are difficult to apply in more common settings with only one KG. Such assumption holds for entities and relations which are popular[14]. For emerging entities better approach that uses information at multiple level, i.e type and instance, can perform better than previous method as shown in Sec. 5.

Explain allows a person to interact with the system to understand a classifier's decision. For a given set of facts, it tries to identify appropriate provenance information, which can be spread across the documents. Approaches likes [9, 7] have been developed for multiple languages but are limited to relations from DBpedia. Hence an approach that takes provenance information from multiple sentences and documents is required.

5 Approach

In order to address the problem, we use complementary approaches, Linguistic Analytic (LA) and Entity Linking (EL), as shown in Figure 1. EL can provide

access to structured information from publicly available knowledge graphs like DBPedia [1], Yago [15], and Freebase [2]. LA can be useful when EL fails to find corresponding entities among other knowledge graphs, especially for emerging entities. Consider Figure 1, where an IE system makes a mistake to classify a partial extracted entity "Broken Glass" as an organization. To determine the credibility of the fact, information from multiple sources with an optional ontology schema can provide complementary information for the classifier.

6 Preliminary Results

In general, facts can be divided into popular and not-so-popular facts. Popular facts involve entities, types, and relations for which relatively more information is available, compared to not-so-popular facts, where there may only be one or a few relevant documents.

As an initial step, I have answered a part of second question which is to identify if the type assertion of emerging entities is correct or incorrect without explanation. For simplicity, we are assuming that all the fact present in the knowledge graph are dubious and need to be verified. A significant number of type error are made by current entity-extraction systems, even with ontologies with limited number of type systems, as in TAC. My current approach takes facts and natural language corpus as input and trains classifiers for each type, combining features surrounding the entities (e.g., Mike Tyson) and representative words from the concept (e.g., Boxer). The following steps are used to learn signal words.

1. For each instance in the training set, retrieve the top-k documents that contains the entity mention and extract surrounding context.
2. Combine surrounding context for all the entities to create a vocabulary set.
3. *Entity Features* (EF): A classifier is trained for each type t, with each word in the vocabulary as a feature and each context is treated as instance. An instance is considered positive if the context is extracted from an entity (e.g., Mike Tyson) belonging to the class and negative when the entity (e.g., Bill Gates) belong to a disjoint class. Perform feature reduction (FR) to select top-n high weighted words as features.
4. *Concept Features* (CF): Combine the documents for each entities of the class. Process document text using LSA to find importance of each words. Perform feature reduction to select the top-n words as features.
5. Combine CF and EF to train a classifier with documents as instances. Assign positive and negative labels following disjoint class axioms. Values for each feature are frequency counts of word present in the document.
6. Test it on the test data.

Dataset: We evaluated the approach on expert engineered gold standard which contained label of the entity, e.g Mike Tyson, followed by the expected type and set of documents with named entity offset. Gold standard contains 2,350 entities of type person, organization and geo-political entity. We used

Approach	AUC
OpenEval	68.90
(EF + FR) + (CF + FR)	**78.15**

Table 1: Area Under Curve (AUC) performance comparison of our approach vs. baseline models (FR is Feature Reduction).

Approach	AUC
EF	76.97
CF	77.06
EF + FR	74.72
CF + FR	77.50
(EF + FR) + (CF + FR)	**78.15**

Table 2: Area Under Curve (AUC) performance of our approach for multiple configurations (FR is Feature Reduction).

Fig. 2: AUC performance results

text corpus of roughly two million LDC documents [6] that included about one million newswire documents, one million Web documents and 99K discussion forum posts, hosted on an elasticsearch [8] instance using a single node and default search settings. To select appropriate number of representative words for each category we experimented with different values of the hyper-parameters of the approach. To avoid biased result we partition the gold standard in size 60%\20%\20% and used 20% to decide the hyper-parameter.

Baseline: We compared our algorithm with a more general state-of-the-art baseline approach, OpenEval [14]. OpenEval is an iterative approach to training classifier for each relation and type. For a given fact, it iteratively uses web search engine to obtain a set of document for training of the classifiers. Poor performing classifiers are improved in next iteration by enhancing the queries. This process is repeated for fixed number of iterations. A similar process is conducted during testing to conclude the credibility of unseen facts.

Discussion: Table 1 shows improvement over state-of-the-art baseline. The improvement achieved by our approach may be explained due to a combination of surrounding words features along with class based keywords features. In order to better understand this, consider our performance at multiple configuration mentioned in Table 2. A minor performance gain is obtained when only surrounding words are considered as features. Relatively better performance is achieved when class level keywords are used as features without feature reduction. However, the best performance is achieved when entity features and class features combined after feature reduction. Hence supporting our second hypothesis.

7 Evaluation

In order to measure if the questions asked in Section 3 are answered, I propose to use evaluation approaches mentioned in [3]. Manually engineered gold standard would suffice to evaluate the performance of confirm phase. However, such an approach might not be good candidate for Narrow Down as it aims to identify a poor quality section of the graph and plan to use ranking. Each domain could be

ranked using expert perception and experience to later compare with the ranking produced by the system. In case of explain I plan to construct a crowd source gold standard expecting the human to assign a score to each of the justification and use it to determine performance of the proposed algorithm.

8 Lessons learned, Open Issues, and Future Work

The main contributions of my PhD dissertation will be to answer the questions in Section 3. To realize the framework, we performed experiments to determine credibility of entity types for emerging entities. Tables 1 and 2 show that combining information at multiple level, i.e type and instance level, yields better classification for errors. One issue is the lack of gold standards on multiple datasets, especially for the Explain. As of now, we have used gold standards available from LDC [6] which follows well-defined annotation guidelines for each entity type and relations. However, creating such guidelines takes time, effort and are expensive. For future work, we plan to extend the existing approach outlined in Section 5 for relations beyond entity types, like 'hasSpouse' or 'holdsPosition'. We plan to develop a methodology to explain errors that are encountered in a knowledge graphs and consider *fluents* where a fact is correct for a time frame.

9 Related Work

There are some previous work on cleaning a KG but is often limited to numerical values, or access ontology schema, or focus on popular entities and leaves a gap to consider a single KG without schema information for emerging entities. Existing approaches to deal with dubious facts can be divided into two categories: *internal* and *external* [12]. *Internal* approaches use the facts mentioned in the knowledge graph. Internal approach like [16] uses information about distributions to identify outliers as incorrect facts. However, the approach is limited to numerical literal values. More general approach [13] models ontological constrains as first order rules to use Probabilistic Soft Logic to infer correct KG from a given KG. However, such approaches are limited to the relations with predefined semantics, e.g., label and mutualExclusion and are not applicable to custom or natural relations like spouseOf.

On the other hand, *external* approaches use resources beside the knowledge graph, such as a Web search engine or access to background knowledge. External approach like [9] uses a popular search engine to help user quickly select correct provenance information for a given fact but covers relations limited to DBpedia. Relatively broader approach is shown in [17] which uses an ensemble of knowledge graphs for the same set of documents with multiple extraction systems to assess the correctness of the facts. However it's unclear how such methods can be applied to single knowledge graphs like NELL and/or to emerging entities. An approach similar to ours is demonstrated in OpenEval [14], which iteratively gathers evidence via Web searches to train classifiers for each relation and type and use them to determine the credibility of unseen facts. Since the approach

relies on Web searches, it works well on popular entities, types, and relations and poorly on emerging entities. We overcome this limitation by using information at multiple levels, i.e type and instance level, to determine of the fact for emerging entities, as described in Sec. 5.

Acknowledgments. I thank to my thesis advisor, Professor Tim Finin.

References

1. Auer, S.e.a.: Dbpedia: A nucleus for a web of open data. In: The Semantic Web. Springer (2007)
2. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: SIGMOD. ACM (2008)
3. Brank, J., Grobelnik, M., Mladenić, D.: A survey of ontology evaluation techniques (2005)
4. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr, E.R., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: AAAI (2010)
5. Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., Zhang, W.: Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In: SIGKDD. ACM (2014)
6. Ellis, J., Getman, J., Fore, D., Kuster, N., Song, Z., Bies, A., Strassel, S.: Overview of linguistic resources for the TAC KBP 2015 evaluations: Methodologies and results. In: TAC KBP Workshop (2015)
7. Gerber, D., Esteves, D., Lehmann, J., Bühmann, L., Usbeck, R., Ngomo, A.C.N., Speck, R.: Defactotemporal and multilingual deep fact validation. Journal of Web Semantics (2015)
8. Gormley, C., Tong, Z.: Elasticsearch: The Definitive Guide. " O'Reilly Media, Inc." (2015)
9. Lehmann, J., Gerber, D., Morsey, M., Ngomo, A.C.N.: Defacto-deep fact validation. In: ISWC (2012)
10. Mayfield, J., McNamee, P., Harmon, C., Finin, T., Lawrie, D.: Kelvin: Extracting knowledge from large text collections. In: AAAI Fall Symposium. AAAI (2014)
11. Mitchell, T., Cohen, W., Hruschka, E., et. al, P.T.: Never-ending learning. In: AAAI (2015)
12. Paulheim, H.: Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic web (2017)
13. Pujara, J., Miao, H., Getoor, L., Cohen, W.: Knowledge graph identification. In: ISWC (2013)
14. Samadi, M., Veloso, M.M., Blum, M.: Openeval: Web information query evaluation. In: AAAI (2013)
15. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A large ontology from wikipedia and wordnet. Journal of Web Semantics (2008)
16. Wienand, D., Paulheim, H.: Detecting incorrect numerical data in DBpedia. In: ESWC (2014)
17. Yu, D., Huang, H.e.a.: The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding. In: COLING (2014)
18. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Semantic Web (2016)