

Generations of Knowledge Graphs: The Crazy Ideas and the Business Impact

Xin Luna Dong
Redmond, Washington
lunadong@gmail.com

ABSTRACT

Knowledge Graphs (KGs) have been used to support a wide range of applications, from web search to personal assistant. In this paper, we describe three generations of knowledge graphs: *entity-based KGs*, which have been supporting general search and question answering (e.g., at Google and Bing); *text-rich KGs*, which have been supporting search and recommendations for products, bio-informatics, etc. (e.g., at Amazon and Alibaba); and the emerging integration of KGs and LLMs, which we call *dual neural KGs*. We describe the characteristics of each generation of KGs, the crazy ideas behind the scenes in constructing such KGs, and the techniques developed over time to enable industry impact. In addition, we use KGs as examples to demonstrate a recipe to evolve research ideas from innovations to production practice, and then to the next level of innovations, to advance both science and business.

PVLDB Reference Format:

Xin Luna Dong. Generations of Knowledge Graphs: The Crazy Ideas and the Business Impact. PVLDB, 16(12): 4130 - 4137, 2023.
doi:10.14778/3611540.3611636

"Science is to test crazy ideas; engineering is to bring these ideas into business."
– Andreas Holzinger

1 INTRODUCTION

Since the birth of modern Knowledge Graphs (KGs) around 2007 (in the same year, Yago [40], DBPedia [4], and Freebase [5] were released)¹, the area has been broadly researched in a multitude of research communities (to name a few, NLP, IR, Data Mining, Databases, Semantic Web). The industry deployment started about a decade ago, when Google launched *Knowledge Panels* in web search in 2012; since then, KGs have been used broadly to support web search (e.g., Google and Bing web search), voice assistants (e.g., Amazon Alexa, Apple Siri, and Google Assistant), and so on, and have made profound business impact.

KGs model the real world in a graph representation, where nodes represent real-world entities or atomic (attribute) values, and edges represent relations between the entities or attributes between entities and atomic values. A piece of knowledge can be considered as a *triple* in the form of (subject, predicate, object), such as (Seattle,

located_at, USA). The *data instances* in a KG follow the *ontology* as the schema, which in itself is represented in a graph form and can be taken as a part of the KG. The ontology describes entity *classes*, often organized in a hierarchical structure and also called *taxonomy*, and meaningful relationships between classes.

KGs can be considered as *semi-structured*: on the one hand, it enjoys clean semantics of structured data powered by the rigidity of schemas (i.e., ontology); on the other hand, it embraces the flexibility of unstructured data by allowing easily adding new classes and relationships. An additional advantage of KGs is that it can seamlessly connect a large number of domains through common entities across domains or relationships between domains (e.g., the *Movie* and *Music* domains can be connected by people who are both actors/actresses and singers, and by the *featured_song* relation). These advantages give KGs a unique position that is both understandable to machines (through ontology) and easy-to-understand by human beings (blessed by the structure), suitable to facilitate understanding in search, question answering (QA), and dialogs, to power recommendation through the graph structure, and to display information for human understanding (in attribute-value pairs), comparison (in tables), and explanation (in paths in the graph).

With the widespread applications of KGs, *how to model and capture all valuable knowledge in the world* has emerged as a prominent research area. This paper delves into this subject through the author's journey in the past decade, enriched with extensive scientific research and production deployment experiences gained at esteemed companies like Google, Amazon, and Meta.

1.1 Generations of knowledge graphs

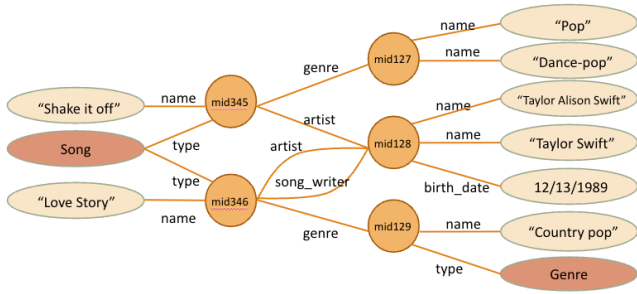
In this paper, we discuss a few generations of KGs. The first generation is *entity-based KGs*, where both ontology and data are more rigorous, and nodes in the graphs are mostly entities that have one-to-one correspondence with real-world entities (see Figure 1(a) as an example). Most well-known generic KGs, such as Yago [40] from academia and Google KG [39] from industry, are entity-based KGs. We discuss this generation in Section 2.

The second generation is *text-rich KGs*, where ontology and data allow much more ambiguities, and nodes in the graphs are more often just free texts. With the text nodes, the graph is mostly in the form of a bipartite graph, as depicted in Figure 1(b). Text-rich KGs are often used to model domains where structure is sparse while ambiguities are abundant, with vague and fluid semantic boundaries between values and even classes, such as *Product*, *Bio-informatics*, and *Health*. Section 3 discusses this generation.

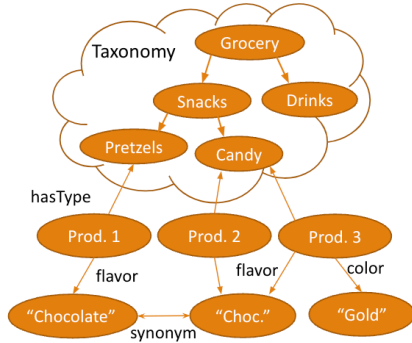
The upcoming generation is not fully shaped yet and we call it *dual neural KGs* for now. It encodes knowledge explicitly as triples (as in KGs) and implicitly as embeddings (as in language models). The same piece of knowledge may co-exist in both forms or stay

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit <https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.
Proceedings of the VLDB Endowment, Vol. 16, No. 12 ISSN 2150-8097.
doi:10.14778/3611540.3611636

¹There are two knowledge bases before all KGs discussed in this paper, Cyc (cyc.com) and WordNet (wordnet.princeton.edu); they are limited in scope and scale because of hand-crafting.



(a) An example entity-based KG in the music domain. Nodes are mostly entities, each with an ID.



(b) An example text-rich KG in the product domain. The top depicts the taxonomy, which can be a rich and deep hierarchy. The bottom depicts data instances, where attribute values are mostly texts; as such, it is mostly in the form of a bipartite graph (except edges like "synonym").

Figure 1: Example knowledge graphs.

on one side that is more suitable, and there is smooth transition between the two forms to allow harmonic blending. Section 4 discusses why we believe *co-existing* is the key for success, at least in the near future.

1.2 The recipe from innovation to practice

KG is an area that has witnessed success both in research and in industry. As we discuss the evolution of KGs, we employ it as an example to illustrate the cycle from innovation to production practice, and subsequently to the next round of innovation. This iterative cycle often comprises several stages, each contributing to impacts from initial to profound.

- (1) **Feasibility:** The cycle first starts with a (or a series of) prototype or an experiment, showing the feasibility of a crazy idea, which sometimes seeds a new field.
- (2) **Quality:** The second stage focuses on gradually improving the quality of the solution (a model, an algorithm) to production quality, which enables trustworthy and pleasant user experiences. This is the key stage to land an innovation as a tangible product: unless attaining production quality, a research idea will only remain research.
- (3) **Repeatability:** Once we achieve success with the initial product, usually within a limited scope (a few domains,

or working under a set of constraint conditions), the next stage is to repeat the success for larger scopes like broader domains. This stage often emphasizes building pipelines to facilitate automation, and employing machine learning (ML) models to minimize manual work. It is a stage leading to much higher business impact.

- (4) **Scalability:** Although repeatability could lead to impact enhancement of 1-2 orders of magnitude, it oftentimes falls short of achieving true scalability, demanding impacts of thousands or millions of times. Scalability often necessitates a new set of solutions that substantially reduce costs and eliminate all manual work from the loop.
- (5) **Ubiquity:** Finally, ubiquity seeks to maximize the scope of applicability, to encompass long-tail use cases, to remove any underlying assumption in the solutions. Pursuing such solutions often triggers a new round of innovations, initiating the next cycle (sometimes even scalability can lead to the next cycle).

This paper interweaves the discussions of the three generations of KGs and the innovation-to-practice-to-innovation cycle. Through the former, we illustrate how development of techniques leads to larger and larger business impact; through the latter, we shed insights on how the pursuit of large business impact sparks new innovations. Finally, we reflect on critical factors for production success in Section 5.

2 ENTITY-BASED KNOWLEDGE GRAPHS

Entity-based KGs are the most popular KGs in both academia (e.g., Yago [40], DBPedia [4], *etc.*) and industry (e.g., Google KG [39], Bing Satori KG [23], Alexa KG [2], WikiData [43]). As early as in 2015, it was reported that Google Knowledge Graph shows for about 25% of all Google search queries ².

There are two characteristics for entity-based KGs. First, the ontology of the KGs is normally manually defined with *clear* semantics, where entity types and relationships have few ambiguities or overlaps. For each domain in the ontology, the numbers of entities and relationships are fairly small and thus manageable for manual definition; for example, Freebase contains 52 entity types and 155 relationships in the *Movie* domain.

Second, most entities in entity-based KG are *named* entities, each corresponding to a real-world entity, such as a person, a university, a movie, a song, and so on. There is rarely overlap between the entities; for example, there are no two persons who are the same, and no two movies that are exactly the same, even if they may share the same name.

2.1 Feasibility and Quality: Knowledge transformation

The seed crazy idea behind entity-based KGs is exactly the idea of modeling the world with entities and relationships. In a sense, that is how human beings understand the world: a child would think about the world as herself, her mom and dad, her friends, her kindergarten, the cartoon she likes, *etc.* Now the question becomes,

²<https://searchengineland.com/googles-knowledge-graph-may-show-14th-search-queries-212962>

how to identify the entities in the world and discover their relationships from available data sources?

Feasibility: Luckily, the idea of modeling the world with entities and relationships is not new: the DBMS (DataBase Management System) uses *ER (Entity Relationship) Diagrams* to visualize the logical structure of the database. Therefore, entities and relationships in KGs can be transformed from structured data such as relational databases. Wikipedia [13], which started in 2001 and describes entities and provides hyperlinks from one entity page to another, conveniently becomes a starting point for collecting knowledge. Wikipedia Infoboxes can be transformed to entities and relationships in a straight-forward way (see an example in https://en.wikipedia.org/wiki/William_Shakespeare); this spurs successful early KGs such as Yago, DBPedia and Freebase.

Quality: The high accuracy of Wikipedia data also guarantees reliability of the derived knowledge. So as far as the transformation is carefully curated to ensure semantics correctness, we can achieve high quality. Since 2012, KGs have been used in production as a trustworthy data source, and Wikipedia has been serving as the major source for the majority of generic KGs even now.

2.2 Repeatability: Knowledge integration

With the success of transforming Wikipedia Infoboxes into knowledge, we naturally wish to enrich knowledge from other structured sources, such as *IMDb* for movies, *MusicBrainz* for music, and *Goodreads* for books. These sources may supplement Wikipedia, oftentimes about torso to long-tail entities (in terms of popularity). However, each of these sources organizes its data in a different way, so the next question becomes, *how to integrate the knowledge transformed from different structured sources?*

The knowledge integration problem is one form of *data integration*, and it needs to resolve three types of heterogeneities [21]:

- **Schema heterogeneity:** Different data sources may express the same entity type and relationship in different ways (e.g., first name and last name vs. full name). *Schema alignment* aligns source schemas with the KG ontology.
- **Entity heterogeneity:** Different data sources may represent the same real-world entity with slightly different names, and provide different attribute values (e.g., Xin Dong from Univ. of Washington vs. Xin Luna Dong from Meta). *Entity linkage* links such entities such that we have a distinct node in the KG to represent a real-world entity. This problem is even more tricky as different entities may share the same name (thus *entity disambiguation*).
- **Value heterogeneity:** Different data sources may provide different attribute values for the same entity, some of which may be imprecise or out-of-date (refer to the same example for entity heterogeneity). *Data fusion* decides among different, and possibly conflicting values, which are correct and up-to-date values.

Among the three problems, schema alignment is mostly done manually to ensure semantics correctness in knowledge transformation; data fusion is less prominent when we restrict knowledge sources to a few authoritative ones. Entity linkage stands out as a critical problem to solve when we link multiple sources, each

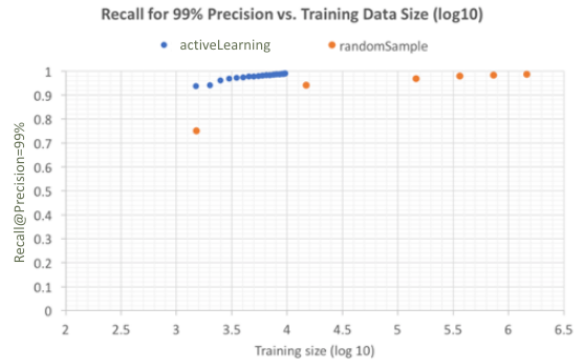


Figure 2: Entity linkage quality with random forest on movies and people between Freebase and IMDb [15]. We are able to achieve over 99% precision and recall with 1.5M labels. When applying active learning to selectively introduce labels, we can achieve the same quality with 10K labels.

of which often has millions of entities or more, making manual linkage implausible.

Entity linkage is a problem with decades of research, dating back to 1969 [22]. In practice, tree-based models have been proved to be effective solutions for entity linkage. Figure 2 shows that we can train random forest models that take attribute-wise value similarities as features, and obtain over 99% precision and recall when linking movies and people between Freebase and IMDb. In addition, the figure shows that although very high precision and recall could require a large number of training labels, applying active learning can reduce training labels by orders of magnitude while maintaining similar linkage quality.

Knowledge integration, especially entity linkage, allows us to repeat the success of knowledge collection from Wikipedia to multiple authoritative structured sources. Most of large KGs harvest data from a variety of sources; for example, Freebase takes data from MusicBrainz, NNDB, Fashion Model Directory, etc..³

2.3 Scalability: Knowledge extraction from semi-structured websites

As discussed in Section 1.2, repeatability does not necessarily lead to scalability. We can transform data from tens of structured sources into knowledge and integrate them to create a holistic KG; however, a lot of manual interference is needed and it is hard to scale up to thousands, or even millions of sources. On the web there are numerous *semi-structured websites* (e.g., rottentomatoes.com), where each page represents a *topic entity*, and different pages display information in key-value pairs at relatively consistent locations across the pages. These websites are typically populated from large structured data sources, thus serve as good data sources to enrich KGs. If we can automatically extract knowledge from these websites, instead of relying on manual knowledge transformation, we will be able to scale up knowledge collection from structured sources on the

³[https://en.wikipedia.org/wiki/Freebase_\(database\)](https://en.wikipedia.org/wiki/Freebase_(database)).

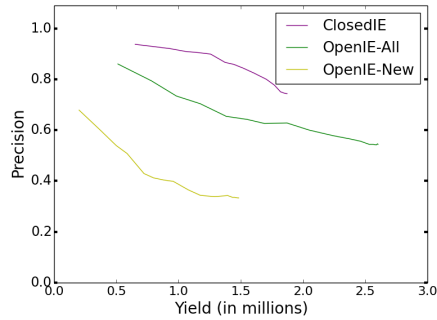


Figure 3: Extraction quality from semi-structured web-sites [34], showing that ClosedIE has achieved over 90% accuracy, whereas OpenIE has shown the promise to increase knowledge, but has much lower accuracy.

web.⁴ Three major techniques has been proposed for knowledge extraction from semi-structured data.

Wrapper induction: Wrapper induction, dating back to 1997, takes manual annotations on a few semi-structured webpages from the same website and induces the extraction patterns expressed in XPath paths that can apply to the whole website [27]. This method works because semi-structured websites are normally populated from underlying databases using some templates (e.g., CSS), and wrapper induction reverse engineers the templates. Although wrapper induction can normally obtain high extraction quality (over 95%), it still requires annotations on every website so is not *truly* web-scale.

Distantly supervised extraction: Distantly supervised extraction started in 2014 [16]; it compares knowledge in existing KGs and data on the semi-structured websites, and generates training data according to the overlaps. Extraction quality is more or less driven by the quality of the training data, so there has been research focused on generating high-quality training data; Ceres [32] does so with careful examination of the structure of semi-structured pages and commonality between pages. OpenCeres [33] further extends this method to annotate (attribute, value) pairs, allowing extracting knowledge for unknown attributes (thus *OpenIE*). This class of methods trains a model per website (precisely, for each cluster of webpages in the website that apply the same template), but the whole process is automatic and thus can scale up to a large number of websites.

GNN-based extraction: The intuition behind GNN-based extraction is that given a semi-structured webpage, one can fairly easily guess what is the topic entity, and what are the attribute-value pairs, without domain knowledge, and even without necessarily understanding the language (e.g., in foreign language). Systems like ZeroShotCeres [34] leverages Graph Neural Network (GNN) to explore both the visual clues and the text semantics, to train one single extraction model for different websites, including even websites in domains where training data do not exist, pushing the boundary of extracting knowledge for *unknown unknowns*.

These three methods are progressively more scalable. As shown in Figure 3, Ceres can achieve over 90% extraction accuracy, thus reaches production quality; whereas extraction for new relations /

domains remains in exploratory stages. Finally, knowledge extraction scales up by automating schema alignment, but there will still be needs for entity linkage and data fusion, where there have been plenty of research for large-scale linkage [25] and fusion [20, 29].

2.4 Ubiquity: Web-scale extraction and fusion

The web is a huge repository of knowledge and it has been the wish of many researchers to extract knowledge from the whole web to achieve ubiquity of knowledge collection. Well-known projects in this line include NELL [10] and Knowledge Vault (KV) [16]. NELL focuses on text extraction, whereas KV extracts knowledge from four types of web contents: texts, semi-structured data (as discussed in 2.3), web tables, and HTML annotations (e.g., according to *schema.org*).

To achieve web-scale, we need an efficient way to generate training data to cover various data patterns. Distant supervision [3, 7, 36, 41] is applied for this purpose, but the training data and thus the extractions are often noisy. Various *knowledge fusion* techniques are proposed to predict correctness of the extractions, such as PRA (Path ranking algorithm) in NELL [10], deep learning based link prediction in KV [16], and graphical models in KV [17]. The graphical models are also used to distinguish extraction errors and source errors, leading to web source trustworthiness evaluation, as in Knowledge-Based Trust [18].

With web-scale knowledge extraction and fusion, NELL extracted 435K knowledge triples and KV extracted 100M triples with over 90% confidence (94M from semi-structured websites). It is orders of magnitude smaller than commercial KGs (to compare, at the same time point Freebase contained 637M triples and Google KG contained 18B triples). Although *web extraction* did not generate a huge volume of knowledge as expected, it led to several important insights. *First*, entity-based knowledge is mainly structured data, so *the best knowledge sources are still structured sources*; thus, knowledge transformation (Section 2.1) and integration (Section 2.2) from well-curated structured sources could be the most effective method to collect high quality knowledge. *Second*, semi-structured websites are major contributors of high-quality knowledge in web extraction, and they can cover long tail knowledge not covered by major structured sources; this insight inspired further investment on knowledge extraction from semi-structured websites, as described in Section 2.3. *Finally*, we find that texts often embrace knowledge not easily captured cleanly by entities, leading to the next generation of KGs, as we will describe in Section 3.

2.5 Summary

To recap, the seed crazy idea behind entity-based KGs is to model the world with entities and relationships, and it faces the challenge that different structured sources express entities and relationships in a heterogeneous way. With knowledge transformation and knowledge integration, major KGs have harvested knowledge from authoritative sources and grown over an order of magnitude over time (e.g., Google KG has grown from 18B triples at launch to over 500B triples⁵). Web extraction from semi-structured websites has also been put in production to supplement long-tail knowledge. Figure 4(a) depicts key techniques as components in building

⁴Webtables [9] is a special form of semi-structured data.

⁵<https://encyclopedia.pub/entry/37713>

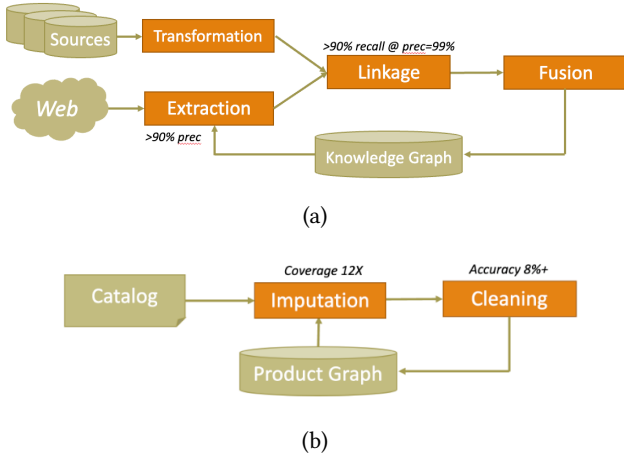


Figure 4: (a) Architecture for constructing an entity-based KG. (b) Architecture for constructing a text-rich KG.

entity-based KGs. Web-scale knowledge extraction has not been as **proliferative** as wished, but has inspired research and technical directions to collect **long-tail knowledge** [14, 28].

3 TEXT-RICH KNOWLEDGE GRAPHS

In many domains like *Products*, *Bioinformatics*, *Health*, *Law*, *Events*, we cannot cleanly model the domain by entities and relationships. We use the *Product* domain as an example to illustrate. First, there can be millions of product types, and many of them are overlapping (e.g., *fashion swimwear* vs. *two-piece swimwear*); thus, defining a clean taxonomy hierarchy is challenging. Second, product attributes are fuzzy and overlapping in nature (e.g., *mocha* vs. *cappuccino* as flavors, where there could be subtle differences but are also often considered as very similar by most customers); thus, entities may not be the best way to capture them. Finally, products are not strictly *named* entities: unlike a person or movie name, product names (e.g., *"Onus 2 Colors Highlighter Stick, Shimmer Cream Powder Waterproof Light Face Cosmetics, creamy Self Sharpening Crayon Stick Highlighter"*) are long, verbose, and concatenation of product type and attributes).

Text-rich KGs are used to model such domains. Instead of setting up clean and strict semantic boundaries between types, relationships, and entities, the *majority* of the nodes in text-rich KGs can be just non-canonical *texts*. Note that different from entity-based KGs, which often also contain text attributes, here text attributes can be dominant, and it is nearly impossible to extract clean entities from these texts. As such, text-rich KGs are more like bipartite graphs rather than regular connected graphs, with topic entities in the domain on one side of the graph, attribute values (or entities) on the other sides of the graph, connected by attributes (see Figure 1(b) as an example).

In the rest of this section, we continue with the *Product* domain as an example to describe the techniques, as the aforementioned challenges are best highlighted in the product domain, and e-business has been prevalent in people's lives. Similar techniques have been applied in other text-rich domains [47].

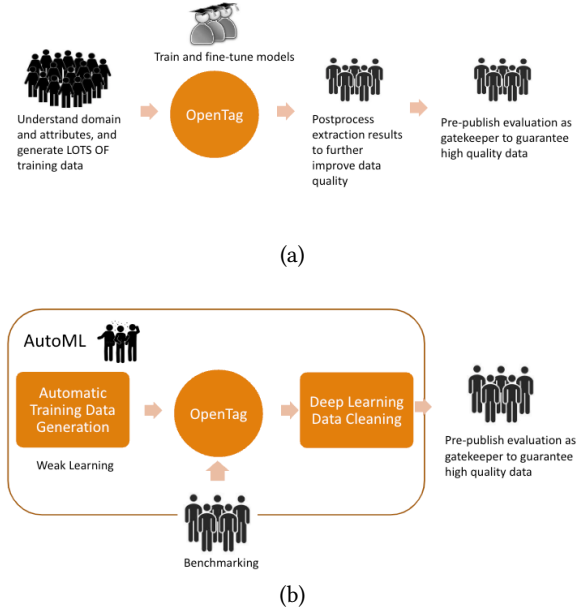


Figure 5: (a) Knowledge extraction pipeline to ensure production quality. (b) Knowledge extraction pipeline with reduced manual work.

3.1 Feasibility: The extraction model

With the huge semantic ambiguities, it is not hard to imagine that in the product domain, structured data are sparse and error prone [47]. The *seed crazy idea* behind text-rich KGs is thus to mine structure and model ambiguity from the structure-sparse source data.

Structure mining relies on knowledge extraction, which requires different techniques from those described in Section 2, since entities are non-named and attributes can be mostly free texts. We resort to product profiles including product names, descriptions, and bullets, and train *Named Entity Recognition (NER)* models to detect patterns that express a particular attribute. Such models, like OpenTag [51], serve as the basis for product knowledge collection.

With the extracted types and attributes, we can mine their relationships (hypernyms, synonyms, etc.) from customer shopping behaviors, such as search, co-view ("customers who viewed this also viewed"), and co-purchase. For example, if users searching for "tea" often buy "green tea", whereas users searching for "green tea" seldom end up buying other types of teas, it hints that "green tea" is a subtype of tea. GNN models have been employed to mine such relationships for types [35] and attribute values [19]. Such methods are also used to establish the substitutes and complements between products [24, 48].

3.2 Quality and Repeatability: The extraction pipeline

Quality: Despite the initial success for NER-based extraction, the quality falls between 85% – 95%, so still mediocre. To achieve production quality (e.g., 90%), a lot of pre- and post-processing is still needed, as shown in Figure 5(a):

- Understand the domain and attributes, and generate training data;
- Fine-tune hyper-parameters to improve the model;
- Postprocess extraction results with rule-based filtering;
- Pre-publish evaluation as a gate-keeper to guarantee high quality results on real data in the wild.

These methods together allow high-quality extraction, often with accuracy above 95%. On the other hand, it introduces a lot of manual work, from labelers, from taxonomists, and from ML engineers and scientists.

Repeatability: To achieve true repeatability for extractions on different attributes and product types, we need a pipeline that is fairly automatic. The following changes, as described in Figure 5(b), aim to remove manual work as much as possible.

- Training data are generated by distant supervision, from existing product Catalog. Since Catalog data could be noisy [19], we still manually label a small number of instances (tens to hundreds) for benchmarking.
- Postprocessing is replaced with deep learning based data cleaning (e.g., transformers, GNNs), leveraging consistency between product descriptions and attributes, between different attribute values of the same product (e.g., snack with sugar in the ingredient is unlikely to be sugar-free), and between products of the same type (e.g., spicy is unlikely to be the flavor of icecreams) [11, 12, 19]. Manual post-processing is only done if ML-based post-processing still cannot achieve the quality bar.
- AutoML pipeline is built to reduce model fine tuning efforts and enable non ML-savvies to tune the models.

With the above improvements, we observed that the time to train and deploy an extraction model can be reduced from a couple of months to a couple of weeks, allowing steadily generating product knowledge to feed e-business features (information display, product comparison, search, recommendation, etc.).

3.3 Scalability: One-size-fits-all solutions

The product domain can contain millions of product types, thousands of product attributes, and hundreds of languages and locales. Even an efficient pipeline as described in Section 3.2 cannot afford to train a model for every combination of product type, attribute, and language. To scale up, we need a solution that is one-size-fits-all.

The product domain is complex because of the huge type variety; even neighboring product types, such as *Coffee* and *Tea*, could have quite different attributes, and different vocabularies and patterns for attribute values. One-size-fits-all models need to be able to understand and leverage the subtle differences between types, attributes, and languages when training the models. Once developed, they would significantly increase the volume of product knowledge and thus the business impact. We next give a few examples.

Multi-type extractions: TXtract [26] deepens its understanding of product types for better extraction in two ways. First, it takes the embedding of the product types as part of the input to the model, so the extraction is type-aware. Second, it employs multi-task learning to predict product types in addition to knowledge extraction, for the model to better understand texts related to type semantics. TXtract shows that it can train one model for 4K product

types, while increasing extraction F-measure by 10% compared to OpenTag as a baseline.

Multi-attribute extractions: The values for different attributes can be more different than values for the same attribute across different product types, thus requiring slightly different models. AdaTag [50], as an example, takes attribute embeddings as input, and applies Mix of Expert (MoE) and HyperNet to leverage the similarities between the attributes (e.g., *flavor* and *scent*, though different, share a lot of common vocabularies) in model training. It can train one model for 32 major attributes whereas still improving quality over training one model per attribute.

3.4 Ubiquity: Extraction from broader sources

Just as collecting entity-based knowledge, we wish to harvest any knowledge existing for the product domain. The techniques described above focus on text information provided by retailers on one single e-business data source, and we can imagine extending knowledge collection in three directions: product images, customer reviews, and multiple e-business websites. These directions can all lead to a new round of innovations, and we briefly describe early results for multi-modal product knowledge extraction.

Product images (both the visual clues and the texts on products) supplement information not existing in product profiles, or enhance information that is vague or ambiguous in profiles. The PAM multi-modal extractor [30] employs a multi-modal transformer to attend across texts and images to improve knowledge extraction; in addition, it uses a generative model, adapted according to the product types, to allow extracting values not observed in training data. Experimental results show that it can improve over text extraction by 11% on F-measure.

3.5 Summary

To recap, the seed crazy idea behind text-rich KGs is to mine structure and model ambiguity for complex domains, and it **faces the challenge that structured data are sparse and noisy**. With one-size-fits-all extraction and cleaning (see Figure 4(b)), Amazon AutoKnow system automatically collected 1B knowledge triples over 11K distinct product types, and considerably extended the ontology and improved Catalog quality [19]. Similar success has been witnessed in other e-business companies [44, 49], and other domains [47].

4 DUAL NEURAL KNOWLEDGE GRAPHS

Comparing with rigorously-structured entity-based KGs, text-rich KGs inject free texts in the structure and thus allow much more flexibility to model complex domains. A natural question to ask next is whether we can completely remove the structure—instead of explicitly modeling knowledge, we capture semantics implicitly such as through embeddings. This question is even more prominent given the recent huge success of LLMs, whose emerging reasoning capabilities [46] seem to have implied failures of *Symbolic AI*⁶. *Will KGs be replaced with LLMs?*

The current: At the current moment, LLMs clearly have not replaced knowledge graphs. First, training an LLM is expensive. As

⁶One application of symbolic AI is *knowledge-based systems*, but here "knowledge" refers to logic rules, instead of structured information as discussed in this paper.

such, it is hard for LLMs to quickly absorb recent knowledge. For example, GPT-4, released in March 2023, is trained with knowledge up to September 2021, with a 1.5-year lag [8]. Second, as broadly known, one major problem for LLMs is [hallucination of non-existing facts](#); our recent study [42] shows that for questions that can be answered using DBPedia data, ChatGPT [1] has a hallucination rate of $\sim 20\%$, and cannot answer $\sim 50\%$ of them. Finally, LLMs can only learn knowledge when it appears often in the training data; as the same study shows, the accuracy in answering questions involving long-tail facts (questions regarding entities in the bottom 33% popularity) drops from $\sim 50\%$ to $\sim 15\%$.

The future: With the above analysis, we envision a KG that encodes knowledge both in the form of knowledge triples and in the form of LLM embeddings, where the former are easier to use for human understanding and explainability, whereas the latter are easier for machine comprehension. We next elaborate with three subsets of knowledge.

- **Taxonomy:** Taxonomy, or the type hierarchies, is what LLMs are good at capturing. With LLMs, it may not be worth explicitly modeling type relationships (e.g., hypernyms, synonyms, etc.), not to mention manually constructing a very deep and complex hierarchy. So tail taxonomy may best reside at the LLM side.
- **Head knowledge:** Training data should be abundant for head knowledge (popular entities and popular attributes) so intuitively there could be a way to teach LLMs head knowledge so they can efficiently address such information needs; in other words, ideally head knowledge reside in both forms. Surprisingly, LLMs can still have a high hallucination rate for head entities (the previously mentioned study shows a hallucination rate of 21% for DBPedia entities with top-33% popularity). One important research problem is how to infuse head knowledge into LLMs to enable precise answers to relevant questions, through model training, or through model fine tuning. Early work in this line includes knowledge infusion [31, 45].
- **Torso-to-tail and recent knowledge:** With the current techniques for LLM training, LLMs are unlikely to be able to effectively incorporate such knowledge, which is lacking in training data. Thus, such knowledge may best reside as triples. Best serving such knowledge requires knowledge-enhanced LLM, which can effectively decide if such knowledge is required for the conversation, seamlessly plug-in external knowledge sources, and do so efficiently. Early work in this line includes knowledge-augmented LLM [6, 37, 38].

How to blend the two forms of knowledge elegantly and how we best address our knowledge needs by leveraging the latest advancements of LLMs remain an open question, and a hot research topic. In addition, how to effectively capture personal knowledge and multi-modal knowledge, leverage them in LLMs to support QA and conversations are even broader research areas.

5 REFLECTIONS: FACTORS TO INDUSTRY SUCCESS

Before we conclude this paper, we reflect on what are key factors to land *crazy science ideas* in industry and lead to real *business impact*.

As observed from a broad range of research directions, there are two necessary conditions. First, the technique has achieved production quality, or, it is *ready*; the bar can be different for different techniques, but high for knowledge correctness, normally 90% to 99%. Second, the technique enables significant scale-ups of productivity, or, it is *essential*. We now illustrate using KG-relevant topics.

Industry successes: A few areas have witnessed prosperity in industry, including (1) knowledge-based QA, which improves the way we address people’s information needs; (2) entity linkage, which is critical in knowledge integration, as discussed in Section 2.2; (3) closed information extraction (ClosedIE), which is critical for scaling up knowledge collection for both entity-based and text-rich KGs, and (4) knowledge cleaning, which is important to filter imprecise knowledge from sources and from extractions. All of these fields satisfy the two conditions: reaching production quality, and increasing productivity significantly to provide better user experiences.

Not-yet successful: On the other hand, there are research areas that have not seen prevalent industry applications.

- (1) Automatic schema alignment: Schema alignment for a few sources is typically done manually by professional taxonomists to ensure 100% correctness (as discussed in Section 2.1), whereas schema alignment at the web scale is done through ClosedIE.
- (2) Knowledge fusion: Integrating knowledge from a few authoritative sources does not encounter too many conflicts, and integrating knowledge from a very large number of sources is not popularly deployed in industry, so the need for fusion is still limited.
- (3) Link prediction (aka, knowledge inference): Link prediction has not achieved the quality to reliably add inferred knowledge into KGs; another use of it, to detect incorrect information, has been incorporated into knowledge cleaning techniques.
- (4) OpenIE: Extracting knowledge where entities and relations are all texts without any restriction has been considered as a promising way to significantly increase the volume of KGs, but the quality has not been satisfactory for production; recent LLMs have allowed much better ways to capture such *fully open* knowledge.

These techniques miss one of the two factors thus have not seen broad production uses; however, we find that they have inspired new research topics, and sometimes those newly developed techniques "replaced" techniques of the original form.

6 CONCLUSIONS

This paper describes generations of knowledge graphs: entity-based KGs, text-rich KGs, and dual neural KGs. In addition, it uses the evolution of KG construction techniques to illustrate the cycle from innovation to production and further to next round of innovation, containing five stages: feasibility, quality, repeatability, scalability, and ubiquity. The recent big success of LLMs shows new directions for knowledge collection and knowledge encoding, which surely will bring knowledge encoding, collection, and mining to the next era and further push business impact to the next level.

ACKNOWLEDGMENTS

Sincere thanks to Alon Halevy, Divesh Srivastava, Gerhard Weikum, and Yifan Ethan Xu for the careful reading of the paper and many helpful suggestions to improve the paper.

REFERENCES

- [1] [n.d.]. ChatGPT. <https://chat.openai.com/>.
- [2] 2018. How Alexa keeps getting smarter.
- [3] E. Agichtein and L. Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *DL*.
- [4] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *Proc. of ISWC*.
- [5] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*. 1247–1250.
- [6] Sebastian Borgeaud, Arthur Mensch, and etc. Jordan Hoffmann†. 2022. Improving language models by retrieving from trillions of tokens. *arXiv* (2022).
- [7] Sergey Brin. 1998. Extracting Patterns and Relations from the World Wide Web. In *Proc. of the WebDB Workshop*.
- [8] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv*.
- [9] Michael J. Cafarella, Alon Y. Halevy, Daisy Zhe Wang, Eugene Wu, and Yang Zhang. 2008. WebTables: exploring the power of tables on the web. In *PVLDB*. 538–549.
- [10] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. Hruschka Jr., and T. Mitchell. 2010. Toward an Architecture for Never-Ending Language Learning. In *AAAI*.
- [11] Kewei Cheng, Xian Li, Zhengyang Wang, Chenwei Zhang, Binxuan Huang, Yifan Ethan Xu, Xin Luna Dong, and Yizhou Sun. 2023. Tab-Cleaner: Weakly Supervised Tabular Data Cleaning via Pre-training for E-commerce Catalog. In *ACL*.
- [12] Kewei Cheng, Xian Li, Yifan Xu, Xin Luna Dong, and Yizhou Sun. 2022. PGE: Robust product graph embedding learning for error detection. In *VLDB*.
- [13] Ludovic Denoyer and Patrick Gallinari. 2006. The Wikipedia XML corpus. *SIGIR Forum* 40, 1 (2006), 64–69.
- [14] Xin Luna Dong. 2016. Leave No Valuable Data Behind: The Crazy Ideas and the Business. In *VLDB*.
- [15] Xin Luna Dong. 2019. Building a broad knowledge graph for products. In *Proc. of ICDE*.
- [16] Xin Luna Dong, Evgeniy Gabrilovich, Jeremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmman, Shaohua Sun, and Wei Zhang. 2014. Knowledge Vault: A Web-scale approach to probabilistic knowledge fusion. In *SIGKDD*.
- [17] Xin Luna Dong, Evgeniy Gabrilovich, Jeremy Heitz, Wilko Horn, Kevin Murphy, Shaohua Sun, and Wei Zhang. 2014. From Data Fusion to Knowledge Fusion. *PVLDB* (2014).
- [18] Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-based trust: estimating the trustworthiness of web sources. In *VLDB*.
- [19] Xin Luna Dong, Xiang He, Andrey Kan, Xian Li, Yan Liang, Jun Ma, Yifan Ethan Xu, Chenwei Zhang, Tong Zhao, Gabriel Blanco Saldana, Saurabh Deshpande, Alexandre Michetti Manduca, Jay Ren, Surender Pal Singh, Fan Xiao, Haw-Shiuan Chang, Giannis Karamanolakis, Yuning Mao, Yaqing Wang, Christos Faloutsos, Andrew McCallum, and Jiawei Han. 2020. AutoKnow: Self-Driving Knowledge Collection for Products of Thousands of Types. In *SigKDD*.
- [20] Xin Luna Dong and Felix Naumann. 2009. Data fusion—Resolving data conflicts for integration. *PVLDB* (2009).
- [21] Xin Luna Dong and Divesh Srivastava. 2013. Big data integration. *PVLDB* (2013).
- [22] Ivan P. Fellegi and Alan B. Sunter. 1969. A Theory for Record Linkage. *Journal of the American Statistical Association* 64, 328 (1969), 1183–1210.
- [23] Yuqing Gao, Jisheng Liang, Benjamin Han, Mohamed Yakout, and Ahmed Mohamed. 2018. Building a Large-Scale, Accurate and Fresh Knowledge Graph. In *Proc. of SIGKDD*.
- [24] Junheng Hao, Tong Zhao, Jin Li, Xin Luna Dong, Christos Faloutsos, Yizhou Sun, and Wei Wang. 2020. P-Companion: A principled framework for diversified complementary product recommendation. In *CIKM*.
- [25] Di Jin, Bunyamin Sisman, Hao Wei, Xin Luna Dong, and Danaï Koutra. 2022. Deep transfer learning for multi-source entity linkage via domain adaptation. In *VLDB*.
- [26] Giannis Karamanolakis, Jun Ma, and Xin Luna Dong. 2020. TXtract: Taxonomy-aware knowledge extraction for thousands of product categories. In *ACL*.
- [27] N. Kushmerick, D. S. Weld, and R. B. Doorenbos. 1997. Wrapper induction for information extraction. In *Proc. of IJCAI*.
- [28] Furing Li, Xin Luna Dong, Anno Lergen, and Yang Li. 2017. Knowledge verification for long tail verticals. In *VLDB*.
- [29] Xian Li, Xin Luna Dong, Kenneth B. Lyons, Weiyi Meng, and Divesh Srivastava. 2015. Scaling up copy detection. In *Proc. of ICDE*.
- [30] Rongmei Lin, Xiang He, Jie Feng, Nasser Zalmout, Yan Liang, Li Xiong, and Xin Luna Dong. 2021. PAM: Understanding product images in cross product category attribute extraction. In *SigKDD*.
- [31] Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S. Yu. 2021. KG-BART: Knowledge Graph-Augmented BART for Generative Commonsense Reasoning. In *AAAI*.
- [32] Colin Lockard, Xin Luna Dong, Arash Einolghozati, and Prashant Shiralkar. 2018. Ceres: Distantly supervised relation extraction from the semi-structured web. In *VLDB*.
- [33] Colin Lockard, Prashant Shiralkar, and Xin Luna Dong. 2019. OpenCeres: When open information extraction meets the semi-structured web. In *NAACL*.
- [34] Colin Lockard, Prashant Shiralkar, Hannaneh Hajishirzi, and Xin Luna Dong. 2020. ZeroShotCeres: Zero-shot relation extraction from semi-structured web-pages. In *ACL*.
- [35] Yuning Mao, Tong Zhao, Andrey Kan, Chenwei Zhang, Xin Luna Dong, Christos Faloutsos, and Jiawei Han. 2020. Octet: Online catalog taxonomy enrichment with self-supervision. In *SigKDD*.
- [36] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL*.
- [37] Reiichi Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. WebGPT: Browser-assisted question-answering with human feedback. *arXiv* (2022).
- [38] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. REPLUG: Retrieval-Augmented Black-Box Language Models. *arXiv* (2023).
- [39] Amit Singhal. 2012. Introducing the Knowledge Graph: Things, Not Strings. Google Official Blog.
- [40] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO - A Core of Semantic Knowledge. In *WWW*.
- [41] Fabian M. Suchanek, Mauro Sozio, and Gerhard Weikum. 2009. SOFIE: A Self-Organizing Framework for Information Extraction. In *WebConf*.
- [42] Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023. How Knowledgeable are Large Language Models?
- [43] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. , 78–85 pages.
- [44] Qifan Wang, Li Yang, Bhargav Kanagal, Sumit Sanghai, D. Sivakumar, Bin Shu, Zac Yu, and Jon Elsas. 2020. Learning to Extract Attribute Value from Product via Question Answering: A Multi-task Approach. In *SigKDD*.
- [45] Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *ACL*.
- [46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.
- [47] Chris Welty, Lora Aroyo, Flip Korn, Sara M. McCarthy, and Shubin Zhao. 2021. Rapid Instance-Level Knowledge Acquisition for Google Maps from Class-Level Common Sense. In *HCOMP*.
- [48] Liqiang Xiao, Jun Ma, Xin Luna Dong, Pascual Martinez-Gomez, Nasser Zalmout, Wei Chen, Tong Zhao, Hao He, and Yaohui Jin. 2021. End-to-end conversational search for online shopping with utterance transfer.
- [49] Huimin Xu, Wenting Wang, Xin Mao, Xinyu Jiang, and Man Lan. 2019. SUOpenTag: Scaling Up Open Tagging from Tens to Thousands: Comprehension Empowered Attribute Value Extraction from Product Title. In *ACL*.
- [50] Jun Yan, Nasser Zalmout, Yan Liang, Christan Grant, Xiang Ren, and Xin Luna Dong. 2021. AdaTag: Multi-attribute value extraction from product profiles with adaptive decoding. In *ACL*.
- [51] Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. 2018. OpenTag: Open attribute value extraction from product profiles. In *SigKDD*.