How New is the (RDF) News?

Assessing Knowledge Graph Completeness over News Feed Entities

Tomer Sagi
tsagi@is.haifa.ac.il
University of Haifa
Haifa, Israel

Yael Wolf
ywolf1@staff.haifa.ac.il
University of Haifa
Haifa, Israel

Katja Hose
khose@cs.aau.dk
Aalborg University
Aalborg, Denmark

ABSTRACT

Linked Open Data and the RDF format have become the premier method of publishing structured data representing entities and facts. Specifically, media organizations, such as the New York Times and the BBC, have embraced Linked Open Data as a way of providing structured access to traditional media content, including articles, images, and video. To ground RDF entities and predicates in existing Linked Open Data sources, dataset curators provide links for some entities to existing general purpose repositories, such as YAGO and DBpedia, using entity extraction and linking tools. However, these state-of-the-art tools rely on the entities to exist in the knowledge base. How much of the information is actually new and thus unable to be grounded is unclear. In this work, we empirically investigate the prevalence of new entities in news feeds with respect to both public and commercial knowledge graphs.

KEYWORDS

Knowledge graph extension; Emerging entities; RSS; News

ACM Reference Format:
Tomer Sagi, Yael Wolf, and Katja Hose. 2019. How New is the (RDF) News?: Assessing Knowledge Graph Completeness over News Feed Entities. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3317702

1 INTRODUCTION

Humanity has been generating text since the dawn of civilization. Embedded within are stories of people, their actions, and experiences. Some are well known, their stories repeated and eventually coded into the canon of human knowledge as structured data. This type of information can be queried and disseminated in a variety of forms, ensuring its availability. However, most people, and everything about them, are lost to the anonymity of unstructured text. Similarly, modern texts, such as news articles, blogs, and web-pages, contain huge amounts of unstructured text, of which but a fraction is harvested and codified into structured forms, such as knowledge graphs.

Knowledge graphs (KG) have evolved on the World Wide Web as a standard for representing knowledge as a network of entities associated with facts and connected to other entities via properties. An increasing variety of tools to manage KG exists under the conceptual umbrella of the Semantic Web. Some KG are hand-built by domain experts, some converted from structured sources (e.g., DBpedia [2], which is created from the structured elements of Wikipedia), and some result from semi-automated extraction performed on unstructured texts, e.g., Dacura [30].

Once a Knowledge Graph is created, keeping it up to date entails monitoring its sources for new facts and emerging entities. The challenge of extracting entities and facts from unstructured text to extend existing knowledge graphs is attracting increasing attention from both industry (e.g., ambiverse.com and heuritech.com) and academic research [25, 29, 30, 34]. However, the need to ground the extracted information in entities that exist in the KG leaves many entities and predicates behind. Consider Knowledge Vault [9] (KV), widely considered state-of-the-art in this realm. KV is trained under a Local Closed World Assumption (LCWA) [9, 12], where training and testing statements are ignored if the statement does not appear in Freebase (its evolving knowledge graph). Thus, KV is trained in the absence of true new information. An empirical evaluation of the LCWA approach in [9] finds that it misses by ten percentage points when tested on data labeled without this assumption. This 10% may very well be the new information uncovered by KV! Furthermore, many of the proposed techniques [20] rely on Wikpedia and derived sources, such as Wikilinks, Wikidata, DBpedia, and YAGO, for grounding. Wikipedia growth has essentially plateaued [37]. Thus, requiring new entities and facts to be grounded in existing KG limits accepted information to high-profile people, places, events, and products on which some knowledge already exists.

Concurrently, Linked Open Data and the RDF format have become the premier method of publishing and representing structured data representing entities and facts. Specifically, media organizations such as the New York Times[1] and the BBC[2] have embraced Linked Open Data as a way of providing structured access to traditional media content, including articles, images, and video. To ground RDF entities and predicates in existing Linked Open Data sources, dataset curators provide links for some entities to existing general purpose repositories such as WordNet and DBpedia. However, much of the information is actually ungrounded, and media organizations do not regularly contribute this ungrounded information back to the general purpose repositories.

Despite efforts specifically targeting news [13], which provide a tool to extract new entities from unstructured RSS news feeds, very little research targets these missing/emerging entities. A recent survey [29] stressed the scarcity of knowledge extension approaches

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317702

[1] https://developer.nytimes.com/
[2] https://www.bbc.co.uk/ontologies

that identify and attempt to complete missing entities and not only facts about existing entities. In the latest round of the DBpedia challenge, where participants are required to locate missing information to be added to DBpedia, the winner (Matteo Cannaviccio, Roma Tre University) professed to be missing a good ground truth for measuring against and reported recall (coverage) of only 23% on a limited set of simple phrases, such as âĂIJ[Person] married [Person]âĂİ.

The limited amount of work targeting knowledge extension and the absence of datasets and evaluation methods for KG extension tools is puzzling. At face value, we cannot rule out the possibility that it indicates the absence of a problem. Perhaps the number of new entities in news streams is very low and established knowledge graphs are doing a good job of collecting them? A hint to the magnitude of the problem can be found in [13]. The authors created a manual golden standard for entity disambiguation by selecting a random 1 % of their diverse RSS feed corpus and from the resulting 70K sentences selecting only 479 entity pairs from more prevalent relation types. On this small sample, which was slightly skewed towards more popular entities by the selection method, only 456 of the 934 entities were found in DBpedia. However, it is unclear, how representative this small sample is with respect to the prevalence of new entities in the whole corpus. Furthermore, additional questions regarding the prevalence and characteristics of new entities and facts in the news remain unanswered.

To answer these questions, in this work we empirically investigate the prevalence of new facts and new entities in news articles with respect to large established public and commercial knowledge graphs. We thereby provide insight on the following questions by evaluating them over articles collected from popular RSS feeds:

(1) What percentage of the entities in news feeds is available in a public knowledge graph close to its publication?
(2) How does this figure change over time?
(3) What types of entities are better captured and which types are less represented and how does this reflect on the prevalent methods of knowledge graph extensions?

The remainder of this paper is structured as follows. In Section 2 we describe the problem of knowledge graph extension and the major techniques in use in this realm. In Section 3 we describe related domains and their relation to this work. Section 4 describes the methods we use and justify some of the methodological choices made. We then present the results of the analyses performed (Section 5), and discuss possible explanations and the implications of these results.

2 BACKGROUND

RDF Knowledge Graphs (KG) have become the de-facto standard for representing structured knowledge on the Web. An increasing variety of tools to manage and extend KG exists under the conceptual umbrella of the Semantic Web [3]. KG are based on entities, representing people, places, and abstract things, such as events. KG also contain facts describing entities through predicates. Predicates can be used to form meaningful relations between entities as well. For example, consider Figure 1, representing a small knowledge graph about two persons married to each other. The two entities *BO* and *MO* are connected via the predicates *sc:givenName* and *sc:familyName* to literals describing them and to each other via predicates using *sc:spouse*. The emergence of web ontologies has allowed further conceptualization of knowledge graphs. RDF Schema (RDFS) and the family of Web Ontology Languages (OWL) provide the ability to recognize groups of entities as a class, create *sameAs* links between entities, create *equivalentClass* links between classes, denote different predicates as *subProperties* of a general property, and additional entailment mechanisms. For example, the entities in the small example described above are instances of class *Person* maintained by schema.org[3].

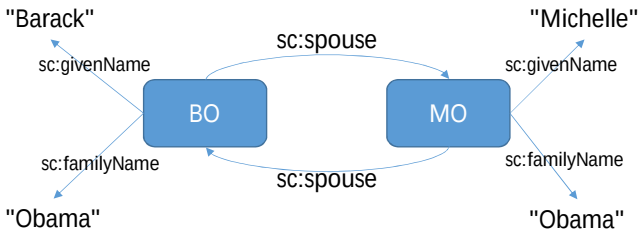

Figure 1: A Small Example Knowledge Graph

These mechanisms have allowed creating one of the most substantial achievements of modern information management, namely, the Linked Open Data (LOD) cloud, a network of KG from a diverse set of domains containing over 201 billion facts[4]. Its existence enables knowledge-based artificial intelligence services such as question answering and reasoning. Such open KG lower the entry barrier for knowledge-based application vendors by giving them access to information in context on a scale previously available only to the largest of corporations, such as Google, Microsoft, and Yahoo. Furthermore, public access to information is currently mediated by search engines that provide mostly links to websites containing the requested information rather than direct answers to questions and information in its context. Large-scale KG representations of both historic and current texts allows unmediated access to information within its factual context rather than within a web page designed by a commercial entity.

Some KG are hand-built by domain experts, some converted from structured sources (e.g., DBpedia), and some result from (semi-)automated extraction processes performed on unstructured texts (e.g., T2KG [19] and Dacura [30]). Of the three methods, the latter (extraction from text) holds the greatest potential. A report for The Economist, Cukier[5] estimated that over 95% of data in existence is unstructured.

The process of extending a knowledge graph by extracting entities and predicates from unstructured text is a complex one. As an example, consider the following article published by the Time Magazine feed[6].

Example 2.1. The Oldest U.S. Military Survivor of the Pearl Harbor Attack Has Died at the Age of 106 Ray Chavez, 106, had been battling pneumonia.

[3]https://schema.org/Person, retrieved February 1st 2019
[4]Derived from https://lod-cloud.net/, retrieved January 31st, 2018
[5]http://www.economist.com/node/15557443, retrieved July 16th, 2018
[6]http://feeds.feedburner.com/time/topstories, retrieved November 22nd, 2018

Figure 2 divides the process into three levels. On the linguistic level, information extraction techniques are used to convert text to entities and predicates (facts). Named Entity Recognition (NER) identifies entities (e.g., U.S. Military, The Pearl Harbor Attack, and Ray Chavez) followed by Coreference Resolution, which identifies other mentions of these entities (e.g., The oldest U.S. Military Survivor referring to Ray Chavez). Relation/predicate extraction attempts to attribute facts to extracted entities or create relations between them (e.g., Ray Chavez has died and is a survivor of The Pearl Harbor Attack).

The factual level has also been termed *Knowledge Graph Extension* where extracted entities are disambiguated against the existing knowledge graph. This process has also been called *Entity Linking*. If the entities are not linked to an existing KG entity with sufficient confidence, the system can attempt to predict whether or not they are new/emerging entities and thus should be inserted into the graph. The process of entity extraction and linking (EEL) is often used to annotate entities within unstructured texts with no intent of extending a knowledge graph. Similarly, extracted predicates between known entities can be predicted by comparing them to known facts involving the associated entity/entities in the KG. The problem of relation predication can be associated with the socially prominent, emerging field of fact checking. With the increasing ubiquity of self publication platforms (e.g., blogs), social networks, and new-media organizations, there is growing concern regarding the validity of facts. Identification of false information portrayed as a fact has become an active research area [7, 10, 28] but is not the focus of this work. The complete process of identifying emerging entities and new predicates was also referred to as *KG completion*.

Although in this work we focus on the factual level, for completeness we describe the conceptual level as well. On the **conceptual** level, *Ontology Learning* processes (e.g., [24]) try to generalize a collection of entities into a conceptualization of the domain. For example, after encountering multiple statements about entities declaring them to be democratic senators, we may deduce that *Democratic Senator* is a class. Similarly, the fact that Democratic senators and people share many traits but that some traits are unique to both classes can be used to derive candidate properties, adding to the evolving ontology. These candidate classes and properties can be aligned with an evolving ontology or with an external ontology using ontology alignment techniques [17].

3 RELATED WORK

Detecting a novel document in a stream of documents [18] is a task related to information retrieval (IR). However, instead of identifying novel entities or facts, existing work rather identifies documents containing novel information. Derczynski et. al. [8] evaluate NER tools on twitter tweets, which are comparable to RSS feed articles in terms of length. However, their evaluation focused on the performance of the tools on this dataset rather than the actual existence of entities in the knowledge graph and how this changes over time.

Open information systems, such as Reverb [11] and OLLIE [26], do not use a fixed ontology to classify extracted entities and predicates, but rather identify entities and facts and extract a common ontology. However, the information extracted is not linked to an existing knowledge graph. Similarly, other approaches identify information in structured sources, such as Wikipedia info boxes or other structured HTML elements (YAGO [16], DBpedia [2]) using a fixed ontology. While these tools do create knowledge graphs, they actually recreate the graph each time they are run rather than extend an existing graph. Conversely, approaches, such as NELL [6], DeepDive [32], and Google's Knowledge Vault [9], attempt to ground the information in existing KG.

All of the above mentioned systems use techniques that have been studied and evaluated separately as well. Information extraction technologies pioneered in the NLP domain, and have recently been increasingly used in conjunction with Semantic Web technologies and use cases [25, 29]. As described in the previous section, entity extraction and linking (EEL) is the process of identifying entities in unstructured text and linking them to existing knowledge bases. A wealth of work on this subject exists; recent advances include features using neural networks [20] and word embeddings [22], which can be tailored to better support events [31]. A common approach is to split this process into two distinct steps, the first being named entity recognition (NER) [1], also known as Mention Detection, and subsequently disambiguating the detected entities against a knowledge base. Approaches to perform the latter are called either entity disambiguation (ED) or entity linking (EL) [38]. However, it is those entities that are not linked/disambiguated that interest us the most in this paper, and these often do not even appear in the golden standard against which these techniques are measured.

Discovering new entities to be added to a KG is a type of knowledge graph refinement. It has also been called emerging entity (EE) discovery or out-of-knowledge-base (OOKB) entity discovery. In a recent survey, under the completeness category, Paulheim [29] has listed the limited amount of work on the subject. Notable work includes that of Hoffart et. al. [15], which attempts to identify emerging entities to be added to a knowledge graph focusing on the case where the names of new entities are ambiguous with existing KG entities. Singh et. al. [33] expand this work by providing a human-in-the loop system to identify and merge these entities to an evolving knowledge graph. Recently, Brambilla et. al. [5] have shown how using seeds of existing people in a curated set of tweets one can identify more examples of emerging entities, specifically people. However, all of these efforts have not provided an empirical analysis of the rate and composition of emerging entities in general and specifically in news streams.

4 METHODOLOGY

4.1 The News

The form in which people keep up with current events has changed over the years. Print as a primary source of news has long since been complemented, and in many ways, replaced with visual media and Internet-based sources. Moreover, even within electronic news consumption, traditional websites are being replaced with social-media based news sharing [21]. For the purposes of this work, we return to RSS feeds, used in previous work on entity extraction from news articles [13]. Although their readership is declining[7] and while most people may receive their news in alternate forms, they

[7] https://www.howtogeek.com/164375/why-google-reader-died-4-alternatives-to-rss-readers/

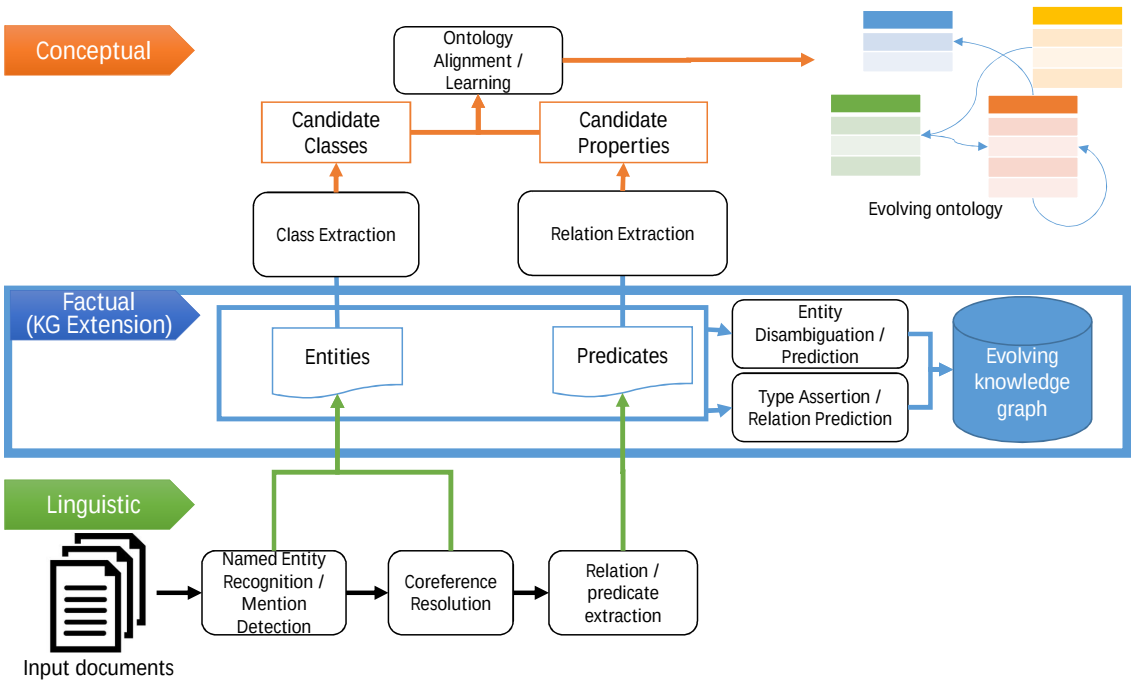

Figure 2: Knowledge extraction process

still represent a reliable method to stream news articles by topic and represent the same news being distributed by other means. Specifically, we took all functioning feeds listed by feeder[8] (one of the most popular RSS feed readers) as the most popular feeds among their users. The 91 feeds cover a wide range of interests and are categorized into the following topics: *News (18), Sports (13), Technology (20), Business (14), Politics (10), and Gaming (16).* Among these, the *News* category is the most diverse, often divided into sub-topics, some of which may overlap with the main topics listed above. We limit ourselves to news articles from English speaking countries since we are comparing against the English language versions of the knowledge graphs, which in general are more up-to-date and extensive than their counterparts in other languages.

4.2 Knowledge Graphs and Entity Linking

In this work, we examine the rate in which new entities appear in up-to-date knowledge graphs (KG) with a publicly accessible API. Table 1 was compiled from previous research [25, 29, 40] and lists known KG considered for this work. Our inclusion criteria were for a KG to be publicly accessible through an API that allows entity linking, and be updated frequently. The knowledge graphs considered are a blend of commercially driven and public. The commercial graphs considered were Google [14], Yahoo[9], and the Microsoft Concept Graph[10]. The public graphs considered were YAGO [36], DBpedia [2] – accessible through Spotlight [27], Freebase [4], Wikidata [39], and Sunflower [23].

Some notable exclusions are: FreeBase, on which a substantial number of entity linking tools (EEL) were tested over the years, was excluded due to the fact that its last public release was in 2012 and that it is now the basis for Google KG. YAGO is rebuilt from its sources using a lengthy process, which is performed infrequently. The YAGO download page[11] dated the latest version on September 20th, 2018 to June 18th, 2017. Wikidata is in many respects equivalent to DBpedia, we therefore chose DBpedia due to ease of use of the DPpedia spotlight API with respect to its Wikidata counterpart Ask Wikidata[12], which is designed to answer natural language questions but can be abused by inputting article text (see discussion on the limited amount of tools for data retrieval from Wikidata by Spitz et. al. [35]).

Our final selection consists of a commercial graph, namely Google KG (GKG) [14] and a public graph namely DBpedia [2]. For both graphs, a public API is provided and both are kept up-to-date through frequent updates from their sources. We employ the DBpedia Spotlight API[13] to parse our articles comprised of the article title concatenated with the article text. Since the Google KG API [14] receives an entity mention and returns a list of matching entities, we perform entity extraction using the Google NLP entity API[14] prior to submitting the mention to the Google KG API.

[8]https://feeder.co/knowledge-base/rss-content/rss-lists-popular-and-useful-rss-feeds/
[9]https://developer.yahoo.com/contentanalysis
[10]https://concept.research.microsoft.com/

[11]https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/downloads/
[12]https://tools.wmflabs.org/bene/ask/, retrieved February 28th, 2019
[13]https://github.com/dbpedia-spotlight/dbpedia-spotlight
[14]https://cloud.google.com/natural-language/docs/analyzing-entities

KG	Commercial?	EL API	Inclusion Decision
DBpedia	No	Via Spotlight	Include
Wikidata	No	Ask Wikidata (partial match)	Exclude
Freebase	No	Unavailable	Exclude, not maintained
Google KG	Yes	Google NLP & KG API	Include
Microsoft Concept Graph	Yes	Limited, class only	Exclude, limited API, no update policy
Sunflower	No	Unavailable	Exclude, not maintained
Wikipedia/Wikilinks/Wikidata	No	SPARQL	Exclude, same source as DBpedia
Yahoo KG	Yes	Rate limited	Exclude - rate limited
YAGO	No	Available	Included in separate evaluation.

Table 1: Knowledge Graphs

4.3 Evaluation procedure

In this section, we detail our evaluation procedure. A companion GitLab repository[15] contains the scripts and tools used to produce the presented results.

4.3.1 Preprocessing. The following preprocessing procedure is performed on all articles. HTML tags and links are removed if present, HTML ASCII codes are converted back to their corresponding utf-8 symbols (e.g. … is converted to …). Article text is truncated at 3200 characters which covers the length of 96.45% of the articles to avoid overloading our human annotators.

4.3.2 Human Annotation. After extracting the entities and looking them up in the KGs we use the *figure eight*[16] crowd sourcing platform for human tagging using the following procedure. A task is generated for each article where the text is annotated using square brackets for the discovered entities. The annotator is tasked with validating that all entities were identified. For example, consider the task shown in Figure 3. An article describing a couple of known local celebrities renovating their home. The entity extraction and linking process against DBpedia correctly identified the Sydney suburb of *Double Bay*, model Jesinta Franklin and the AFL (Australian Football League). However, Buddy Franklin, the AFL player, was misidentified as Benjamin Franklin, the American polymath. After marking "Missing/wrong entities...", a more detailed list of options appears allowing the annotator to list the missing entities and the wrongly identified entities. Tasks flagged for missing entities were manually verified by using the GKG and DBpedia public search interfaces. If they were indeed missing, they were added to a list of missing entities. In the example presented, Buddy Franklin actually exists in DBpedia (and GKG) and was just improperly disambiguated by DBpedia spotlight.

The lookup process and tagging procedure are repeated after one month, and two months to examine changes over time, specifically, which novel entities were added to the knowledge graphs and which remain excluded.

5 RESULTS

A total of 866 cases of missing entities were found out of 13,456 named entities detected by the named entity recognizers. Of these, 378 were missing from DBpedia, 488 were missing from GKG. Figure 4 details the overlap between these groups. For example, of the 378 missing from DBpedia, 81 were missing only from DBpedia and existed in GKG. Similarly, Of the 488 reported above, 191 were missing only from GKG. Note that 297 entities were missing from both DBpedia and GKG, and therefore counted twice in the total number reported above.

After removing duplicates, 408 unique missing entities were left. We classified the entities into six classes according to *schema.org*[17] top level classes, namely People, Products, Creative Works, Organizations, Events and Health & Medical. Table 2 presents the number of entities missing by class with a few demonstrative examples of each class.

To better understand the reasons of omission, we further refined two of the classes described above, namely People and Creative Works. In the People category, missing entities were distributed as follows: 43 journalists, 21 criminals, 17 victims, 20 artists, 31 functionaries in companies, eight sports people, five relatives of famous people, four public functionaries, four researchers, three suspects, three fictional characters, and 37 "others". We speculate that the reason most people are missing from a graph can be attributed to the methods employed by EEL in KG extension, i.e linking new elements to existing ones. Thus, random people (such as the subcategory "others") who appear in a news article will probably not be added to a KG. For example, victims (and criminals) of isolated attacks are less likely to be added than a mass shooter. Having said that, we observed that GKG did contain some entities of victims and criminals of such attacks. The second category, Creative Works was distributed as follows: 33 games, 16 art pieces, three web content, one website, and one "other". Among the missing games, there were old ones, new games released recently, new releases in a series, and games with an upcoming release date which may be the reason why they were not in a KG yet. The art pieces contained mostly characters of animated films that may have a limited number of followers in monitored sources such as Wikipedia, preferring their own wiki-type constructs such as Fandom/Wikia[18] .

In addition, we noticed an interesting trend regarding products. Some companies' products did not appear in one of the KG, mainly Apple and Google products, which were only in DBpedia. This

[15] https://gitlab.com/ts.tomersagi/newnews
[16] https://figure-eight.com
[17] https://schema.org/docs/schemas.html, retrieved Jan. 15th 2019
[18] https://www.wikia.com/explore

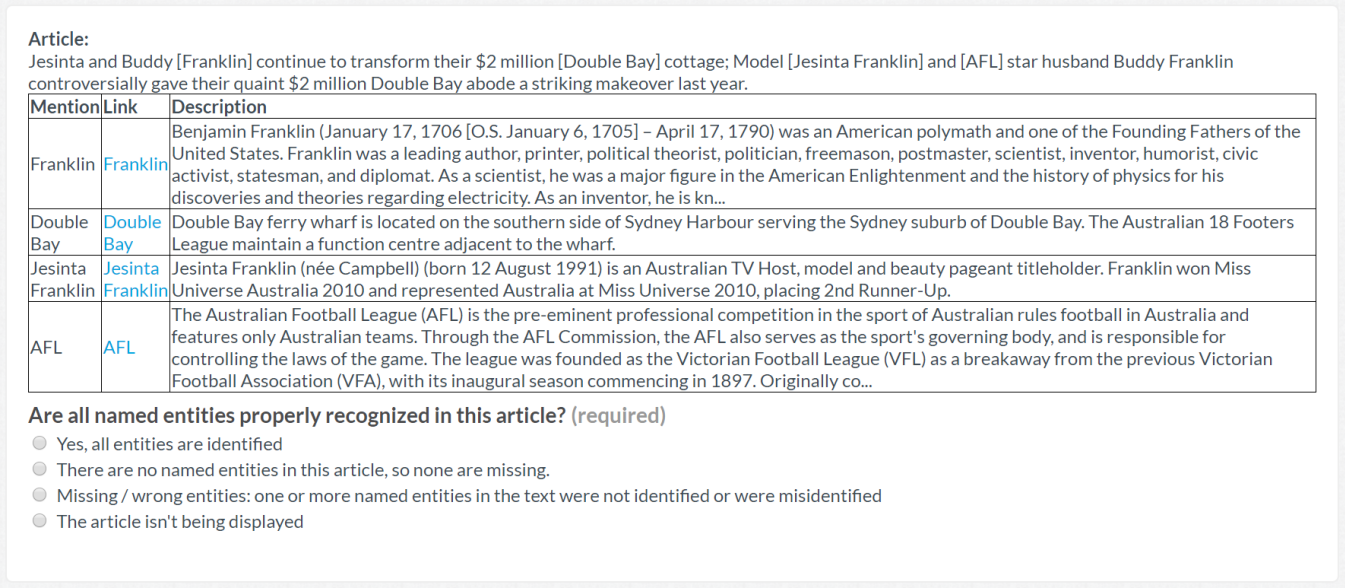

Figure 3: Example of Figure Eight Task

Class	Number	Examples
People	196	Josh Edelson, a photographer and Yasha Haddaji, Nintendo RussiaâĂŹs General Manager, who are both missing from GKG and DBpedia
Products	109	Recon 200, a new headset by Turtle Beach, missing from both GKG and DBpedia and PlayStation 4, video game console, missing from GKG
Creative Works	54	"A Head Of A Young Man ", a painting by Peter Paul Rubens, which is missing from DBpedia, and AssassinâĂŹs Creed Rebellion, a game, which was missing from both GKG and DBpedia
Organizations	41	Chime Banking and Empower, both banks missing from both GKG and DBpedia
Events	7	Assembly Elections in India which was missing from GKG and FortniteâĂŹs Winter Royale Online Tournament, which was missing from both GKG and DBpedia
Health & Medical	1	Acute Flaccid Myelitis, a rare disease, missing from GKG.

Table 2: Classes and Examples

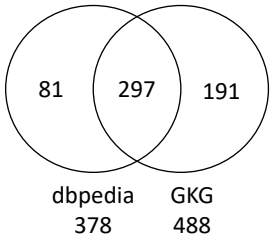

Figure 4: Venn Diagram Depicting Missing Entities

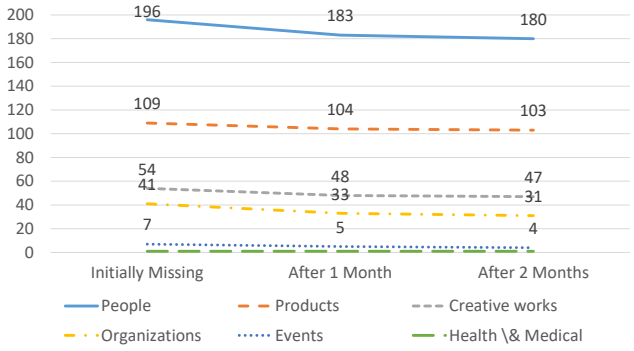

Figure 5: Missing Entities over Time (by class)

could result from company policy but the real reason is unknown to us.

After one month, we reexamined the above unique entities to see which were added to the knowledge graphs. A total of 34 additions were found: eight to DBpedia and 26 to GKG. Six creative works were added: four of them were added to GKG – all video games, three are new, and one has a future release date (February

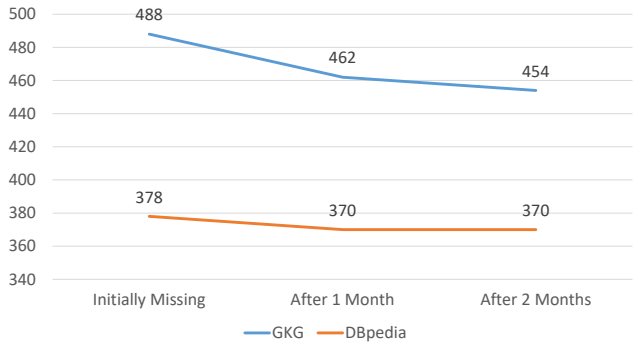

Figure 6: Missing Entities over Time (by graph)

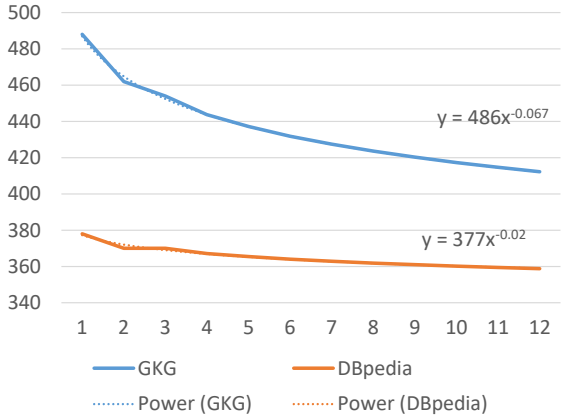

Figure 7: Extrapolated Reclamation Rate

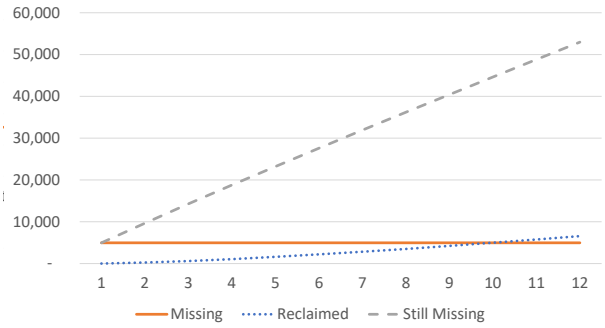

Figure 8: Cumulative Effects of Reclamation Rate

2019) and is part of a series; two were added to DBpedia, the latter game and a fictional character in an animated film. 13 people were added: two to both knowledge graphs, one of which is a young basketball player; seven to GKG including three victims; two to DBpedia, a young soccer player and an up-and-coming actress. Eight organizations were added: one to both knowledge graphs, a financial services company; six to GKG, including a primary school in London. Five products were added: one product was added to both knowledge graphs, and the rest to GKG. Two events were added to GKG, both wildfire events in the United States. After an additional month, we repeated the procedure. Only eight entities were added. All were added to GKG: Three people – young soccer players, two organizations, one event, one creative work and one product. Figures 5 and 6 show the number of missing entities over time by class and graph respectively.

To better understand which types of news streams the missing entities come from, we return to the original feed classification. As mentioned above, feeds were classified at the source into one of five categories, namely Gaming, News, Politics, Sport, and Technology. We manually validated this voluntary classification and found a few articles classified as News but that better fit one of the other categories. Table 3 presents an analysis of the base rate of missing entities in different categories. As is evident, the domains of Gaming and Technology feature a higher base rate than News and Sports. While it is tempting to conclude that there are no new entities in Politics, the small number of articles and the fact that our examination period was between election cycles preclude such a conclusion.

6 CONCLUSIONS AND FUTURE WORK

So... how new is the news? Our empirical analysis found that on average, about six percent (6%) of entities mentioned in a variety of news feeds were missing from the knowledge graph. While the rate may sound small, one must consider that only a small percentage of these may eventually be recovered. To assess the extent of eventual recovery we extrapolated (Figure 7) the number of recovered entities to 12 months by fitting inverse power functions to the original recovery rates. The fitted functions are displayed next to the extrapolated figures. Both functions represent a decaying rate of recovery which best fits the observed data ($R^2 > 0.86$ for DBpedia and $R^2 > 0.98$ for GKG). The overall number of reclaimed entities is projected to be 76 for GKG and 19 for DBpedia, representing 15% and 5% of the missing entities respectively. Figure 8 presents the result of applying this rate and extrapolating the number of newly missing entities to a 12 month period. Under a conservative assumption of a stable rate of new missing entities, it is obvious that the rate of reclamation is vastly out-paced by the rate of new entities. Thus, knowledge about an increasing number of people, products, and creative works is lost over time.

As mentioned above, there is a scarcity of datasets and evaluation methods for identifying and linking novel entities. The work presented in this paper highlights the need for both a dynamic and a static evaluation method for knowledge graph extension. A dynamic method should challenge submitted tools with new information that cannot be found in public knowledge graphs and the static dataset and evaluation protocol should use news articles collected recently with a strictly enforced temporal separation between this evaluation dataset and the data sources on which evaluated systems rely on.

In future work, we intend to explore the underlying assumptions of existing emerging entity identification systems to uncover the methodological reasons for this growing problem. Furthermore, we wish to validate our extrapolation assumptions by scaling this work over time and into additional types of news sources.

Feed Category	Cases of Missing Entities	Total Articles	Missing entity/article	Extracted entities	Missing entities/extracted
Gaming	224	214	1.0	2,786	8%
News	251	557	0.5	4,566	5%
Politics	0	14	0.0	113	0%
Sport	50	252	0.2	2,759	2%
Technology	341	309	1.1	3,232	11%

Table 3: Analysis of missing entity base rate

REFERENCES

[1] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *27th Int. Conf. on Comp. Linguistics, COLING '18*. ACL, 1638–1649.

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In *ISWC/ASWC*. Springer, Busan, Korea, 722–735.

[3] Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific american* 284, 5 (2001), 34–43.

[4] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase. In *SIGMOD '08*. ACM Press, New York, New York, USA, 1247.

[5] Marco Brambilla, Stefano Ceri, Emanuele Della Valle, Riccardo Volonterio, and Felix Xavier Acero Salazar. 2017. Extracting Emerging Knowledge from Social Media. In *WWW '17*. ACM Press, New York, New York, USA, 795–804.

[6] Andrew Carlson, Justin Betteridge, and Bryan Kisiel. 2010. Toward an Architecture for Never-Ending Language Learning. In *AAAI '10*. AAAI Press, 1306–1313.

[7] Sylvie Cazalens, Philippe Lamarre, Julien Leblay, Ioana Manolescu, and Xavier Tannier. 2018. A Content Management Perspective on Fact-Checking. In *WWW '18*. ACM, 565–574.

[8] Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke Van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Inf. Processing and Mgt.* 51, 2 (2015), 32–49.

[9] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In *KDD '14*. ACM, New York, NY, USA, 601–610.

[10] Gonenc Ercan, Shady Elbassuoni, and Katja Hose. 2019. Retrieving Textual Evidence for Knowledge Graph Facts. In *ESWC*. Springer.

[11] Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *EMNLP '11*. ACL, 1535–1545.

[12] Luis Galárraga, Christina Teflioudi, Katja Hose, and Fabian Suchanek. 2013. AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In *WWW '13*. ACM Press, New York, New York, USA, 413–422.

[13] Daniel Gerber, Sebastian Hellmann, Lorenz Bühmann, Tommaso Soru, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. 2013. Real-Time RDF Extraction from Unstructured Data Streams. In *ISWC '13*, Vol. 8218. Springer, 135–150.

[14] Google. 2001. Google Knowledge Graph Search API. Retrieved Feb. 1st, 2019 from https://developers.google.com/knowledge-graph/

[15] Johannes Hoffart, Yasemin Altun, and Gerhard Weikum. 2014. Discovering emerging entities with ambiguous names. In *WWW '14*. ACM Press, New York, New York, USA, 385–396.

[16] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. In *IJCAI '13*. IJCAI/AAAI, 3161–3165.

[17] Prateek Jain, Pascal Hitzler, Amit P Sheth, Kunal Verma, and Peter Z Yeh. 2010. Ontology alignment for linked open data. In *ISWC '10*. Springer, Springer, 402–417.

[18] Margarita Karkali, François Rousseau, Alexandros Ntoulas, and Michalis Vazirgiannis. 2013. Efficient online novelty detection in news streams. In *Web Inf. Sys. Eng. âĂŞ WISE '13. LNCS.*, Lin X., Manolopoulos Y., Srivastava D., and Huang G. (Eds.), Vol. 8180. Springer, Berlin, Heidelberg, 57–71.

[19] Natthawut Kertkeidkachorn and Ryutaro Ichise. 2017. T2KG : An End-to-End System for Creating Knowledge Graph from Unstructured Text. In *Proceedings of AAAI Workshop on Knowledge-based Techniques for Problem Solving and Reasoning*. AAAI Press, 743–749.

[20] Nikolaos Kolitsas, Eth Zürich, Octavian-Eugen Ganea, and Thomas Hofmann. October 31 - November 1, 2018. End-to-End Neural Entity Linking. In *CoNLL '18*. Association for Computational Linguistics, Brussels, Belgium, 519–529.

[21] Anna Sophie Kümpel, Veronika Karnowski, and Till Keyling. 2015. News Sharing in Social Media: A Review of Current Research on News Sharing Users, Content, and Networks. *Social Media + Society* 1, 2 (sep 2015), 205630511561014.

[22] Phong Le and Ivan Titov. 2018. Improving Entity Linking by Modeling Latent Relations between Mentions. In *the 56th Annual Meeting of ACL '18*. ACL, 1595–1604.

[23] Marek Lipczak, Arash Koushkestani, and Evangelos Milios. 2014. Tulip: Lightweight Entity Recognition and Disambiguation Using Wikipedia-Based Topic Centroids. In *1st Int. Work. on Entity Recognition & Disambiguation - ERD '14*. ACM New York, NY, USA, Gold Coast, Queensland, Australia, 31–36.

[24] Alexander Maedche and Steffen Staab. 2001. Ontology learning for the semantic web. *IEEE Intelligent systems* 16, 2 (2001), 72–79.

[25] Jose L Martinez-Rodriguez, Aidan Hogan, and Ivan Lopez-Arevalo. 2018. Information Extraction meets the Semantic Web: A Survey. *Semantic Web* Pre-press, Pre-press (2018), 1—-81.

[26] Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. 2012. Open Language Learning for Information Extraction. In *EMNLP and CoNLL 12', Jeju Island, Korea*. ACL, 523–534.

[27] Pablo N Mendes, Max Jakob, Andrés García-silva, and Christian Bizer. 2011. DBpedia Spotlight : Shedding Light on the Web of Documents. In *7th Int. Conf. on Semantic Sys. (I-Semantics).*, Vol. 95. ACM, 1–8.

[28] Steffen Metzger, Shady Elbassuoni, Katja Hose, and Ralf Schenkel. 2011. S3K: seeking statement-supporting top-K witnesses. In *CIKM*. ACM, 37–46.

[29] Heiko Paulheim. 2016. Knowledge graph refinement: A Survey of Approaches and Evaluation Methods. *Semantic Web* 0 (2016), 1–23.

[30] Peter N Peregrine, Rob Brennan, Thomas Currie, Kevin Feeney, Pieter François, Peter Turchin, and Harvey Whitehouse. 2018. Dacura: A new solution to data harvesting and knowledge extraction for the historical sciences. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 51, 3 (2018), 165–174.

[31] Vinay Setty and Katja Hose. 2018. Event2Vec: Neural Embeddings for News Events. In *SIGIR*. ACM, 1013–1016.

[32] Jaeho Shin, Sen Wu, Feiran Wang, Christopher De Sa, Ce Zhang, and Christopher Ré. 2015. Incremental knowledge base construction using DeepDive. *pVLDB* 8, 11 (jul 2015), 1310–1321.

[33] Jaspreet Singh, Johannes Hoffart, and Avishek Anand. 2016. Discovering Entities with Just a Little Help from You. In *CIKM '16*. ACM Press, New York, New York, USA, 1331–1340.

[34] Anton Södergren and Pierre Nugues. 2017. A Multilingual Entity Linker Using PageRank and Semantic Graphs. In *the 21st Nordic Conf. on Comp. Linguistics, NODALIDA '17, Gothenburg, Sweden*. ACL, 87–95. http://www.ep.liu.se/ecp/131/011/ecp17131011.pdf

[35] Andreas Spitz, Vaibhav Dixit, Ludwig Richter, Michael Gertz, and Johanna Geiß. 2016. State of the union: A data consumer's perspective on wikidata and its properties for the classification and resolution of entities. In *10th AAAI Conf. on Web and Social Media*. 88–95.

[36] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: a core of semantic knowledge. In *WWW '17*. ACM New York, NY, USA, Banff, AB, Canada, 697–706.

[37] Bongwon Suh, Gregorio Convertino, Ed Chi, and Peter Pirolli. 2009. The Singularity is Not Near: Slowing Growth of Wikipedia. In *In the 5th Int. Symp. on Wikis and Open Collab. SE - WikiSym '09*. ACM, 1–10.

[38] Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. 2015. Modeling Mention, Context and Entity with Neural Networks for Entity Disambiguation.. In *IJCAI '15*. AAAI Press, 1333–1339.

[39] Denny VrandeÄŒiÄĆ and Markus KrÃűtzsch. 2014. Wikidata: A Free Collaborative Knowledge Base. *Commun. ACM* 57 (2014), 78–85.

[40] Gerhard Weikum, Johannes Hoffart, and Fabian Suchanek. 2016. Ten Years of Knowledge Harvesting : Lessons and Challenges. *IEEE Data Eng. Bull.* 39, 3 (2016), 41—-50.