Knowledge Graphs: Venturing Out into the Wild

Gerard de Melo(✉)

Rutgers University, New Brunswick, NJ, USA
gdm@demelo.org
http://gerard.demelo.org

Abstract. While we now have vast collections of knowledge at our disposal, it appears that our systems often need further kinds of knowledge that are still missing in most knowledge graphs. This paper argues that we need keep moving further beyond simple collections of encyclopedic facts. Three key directions are (1) aiming at more tightly integrated knowledge, (2) distilling knowledge from text and other unstructured data, and (3) moving towards cognitive and neural approaches to better exploit the available knowledge in intelligent applications.

Keywords: Knowledge graphs · Information extraction · Neural methods

1 Introduction

In the past decade, knowledge graphs have grown from niche academic endeavours to becoming crucial assets for many IT companies. Well-known examples include DBpedia [13], YAGO [11], the Google Knowledge Graph, and Microsoft's Satori. Yet, although we now have vast repositories of facts at our disposal, it appears that our systems often need further kinds of knowledge that are still missing in most knowledge graphs.

This paper surveys three key directions to address the shortcomings of current large-scale knowledge graphs, suggesting paths for moving further beyond simple collections of encyclopedic facts. Section 2 focuses on better knowledge integration for structured data. Section 3 discusses how to connect structured data to the vast amounts of knowledge effectively locked away in unstructured sources. Finally, Sect. 4 proposes cognitive and neural approaches as a means of making better use of such knowledge.

2 Knowledge Integration for Structured Data

In the past, the knowledge acquisition bottleneck was often cited as a key challenge for artificial intelligence. Nowadays, there is a deluge of new sources of machine-readable knowledge. These include not only the RDF-based ones in the Linked Data cloud, but also thousands of open datasets stored in various other formats, and millions of web pages that incorporate structured data.

M. van Erp et al. (Eds.): ISWC 2016 Workshops, LNCS 10579, pp. 1–9, 2017.
https://doi.org/10.1007/978-3-319-68723-0_1

While this abundance of different sources is certainly a blessing, it also brings a set of challenges in downstream applications wishing to make use of such data. What we have at our disposition is in several respects like a rich library with thousands of books. While this library may ultimately be able to serve our information needs, it is not always trivial to find relevant books and locate the desired facts within them. A very early pioneering attempt at addressing this, going even beyond individual libraries, was made in 1895 by Paul Otlet and Henri Lafontaine in their Répertoire Bibliographique Universel (RBU). This universal index would eventually grow to over 15 million index cards, aspiring to systematically organize much of the world's knowledge.

In the digital age, we need tools and algorithms that provide a similar level of universal knowledge organization, yielding pertinent data for a given information need. Converting the various input data formats to a common form such as RDF is just the first step. A more significant challenge is overcoming the heterogeneity of their data models and their incongruent forms of knowledge organization.

One aspect is connecting entities across datasets. The simplest case is when there are shared identifiers. For instance, many resources are linked to Wikipedia or DBpedia for general entities, Lexvo.org [19] for linguistic entities, and WordNet for sense identifiers. In general, however, creating links remains very challenging, despite the long history of work on this. This is particularly true when we aim at entity matching not just between two sources but across a large range of datasets, as this is best done jointly so as to exploit the mutual influence between various candidate matches. The LINDA approach [3] addresses this via a scalable greedy approach that first establishes those links that appear to be easy and reliable. Information about these accepted links is then used to update our beliefs about the accuracy of other potential links. Further algorithms allow us to check for the consistency of entity match links [18]. Another little-studied but important problem is the issue of varying levels of granularity of concepts [21]. Even an entity name such as "London" may refer to multiple competing notions of the entity, e.g., the small City of London, the London metropolitan area, Greater London, the Greater London Built-up Area, or others that may extend as far as to include London Gatwick airport, in addition to various historic definitions. Establishing entity-level links allows us to connect various resources in a cloud of linked resources, similar to the general Linked Data cloud, but possibly also for specific domains as in the Linguistic Linked Data cloud [17].

Even with such links, however, the knowledge is not fully integrated. We have developed algorithms that take a series of separate knowledge graphs as input and produce a single coherent taxonomy, based on ontological principles [1,26].

Another important step is to connect the various properties that are in use across different datasets. To this end, the FrameBase project provides a large schema [28] based on verbs in the English language. This schema draws on the FrameNet lexical resource, extended with additional entries from WordNet for greater coverage. Within the project, a number of heuristics have been developed to automatically connect other ontologies and vocabularies to FrameBase [31]. In some cases, however, manual modeling may be necessary to extend FrameBase to cover more specific properties that cannot straightforwardly be aligned with FrameBase via a 1-to-1 mapping [30]. Hence, we have also developed a user interface that facilitates engaging human experts to define more complex mappings [32].

Examples of integrated knowledge graphs include Lexvo.org [19], which describes languages, scripts, words, and other language-related units, the Universal Wordnet (UWN) [25], which provides multilingual word meanings and their relationships, and MENTA [26], a multilingual taxonomy coherently combining over 200 language editions of Wikipedia. Open challenges include how to cope with incompatible licenses. For instance, the Open PHACTS portal provides data from different sources with incompatible licenses, some of which do not permit derivatives.

3 Connecting Unstructured Data

While information systems excel at processing structured data, large amounts of the world's knowledge are only available via other modalities.

3.1 Text and Language

For natural language text, suitable methods are needed for analysis and knowledge extraction. Standard forms of information extraction (IE) consider only a narrow, predefined set of relations. Although there has been significant progress in this area, including drawing on Web-scale data [36], relying on entire knowledge graphs as seed data [39], and using deep learning models [42], their success often hinges on the availability of relevant training or seed data for each relation.

Open information extraction is a well-known alternative, aiming to cover arbitrary relationships encountered in a text. This open-ended approach may support a broader range of applications. For instance, the PEAK system [46] shows how this allows us to automatically evaluate the quality of a textual summary, given reference summaries. While measures such as ROUGE are often used to automatically evaluate text summarization systems, ROUGE only works reliably when averaging across numerous different texts to be summarized. Often, however, we only wish to evaluate a single summary. This might be a student-written one, for instance, used as a means of assessing reading comprehension. PEAK fills this gap using open IE: Subject-predicate-object triples are used to discover salient units of content expressed in a summary and then such units can be compared between a student-written summary and high-quality reference summaries to automatically assess the student's reading comprehension.

Still, open IE is perhaps best used only as an internal component of knowledge-driven systems. Although the extractions are very useful in certain tasks, they are not sufficiently clean and normalized to be similar to what one encounters in curated knowledge graphs. Additionally, they also mostly neglect n-ary relations (for $n > 2$).

Instead, it may be desirable to obtain extractions with a more well-defined target representation for the extracted knowledge. In particular, is often beneficial for such representations to be compatible with those used in structured knowledge graphs. There is a growing number of knowledge graphs capturing linguistic information. Examples include the aforementioned Lexvo.org [19] and Universal Wordnet projects [25,26]. However, one needs a wide-coverage schema that also covers most kinds of relationships that one expects to extract from text. To this end, the aforementioned FrameBase schema is a suitable target [28], as it brings English verbs and their arguments into the realm of Linked Data. By drawing on interlinked resources such as the Universal WordNet [25], languages other than English are also connected to it. Third-party tools such as the PIKES [6] and KNEWS [2] systems can take us from raw text to extractions based on the FrameBase schema.

Apart from improving the overall accuracy of such systems, ongoing research is focusing on coping with the various intricacies of natural language. Particular phenomena that are being worked on include ambiguity [38], metaphor [33], comparisons [40], nominalizations [8,27], and abstract events [29]. Of particular importance to knowledge-driven applications is the status of clauses and phrases. When a text discusses the “dismissal of the Ambassador”, then in some, but not all contexts, we can conclude that the Ambassador has been dismissed. If a text states that everyone “dislikes that the company is releasing new product X” then the machine should be able to infer that they are indeed releasing X, whereas if it states that they “deny that the company is releasing new product X”, then it is not clear. Similarly, “refusing to secure a loan” is quite different from “managing to secure a loan”. Although these phrases seem trivial for humans to interpret, they differ merely in individual words, and information extraction systems ought to be able to make sense of these differences. We have developed a prototype system that achieves this [23].

Finally, apart from going from text to knowledge, there are also further tasks at the intersection of language and knowledge. An obvious one is to consider the inverse direction of going from knowledge to text, i.e., text generation [43]. Another important task is to make knowledge searchable, i.e., retrieving facts as answers to a natural language query [14,22]. All of these tasks relate to the current trend of developing intelligent conversational agents that rely on natural language skills as well as on knowledge.

3.2 Multimodal Knowledge

In recent years, computer vision has made significant progress and multimodal data has become more connected to language and knowledge. Large cultural heritage collections have become available as Linked Data. Standard computer vision datasets such as ImageNet and Visual Genome are connected to WordNet. Moreover, natural language captions can now be generated automatically for both images and video, by combining deep convolutional neural networks with recurrent models, perhaps incorporating ideas such as multi-faceted attention [15].

Ongoing research is targeting how to go beyond object detection and classification to gain a more complete and thorough understanding of what is going on in an image or video. One direction is to understand images at a higher level of abstraction by predicting not just the concrete objects that they portray, but also the overall activity [9]. This is challenging, because an activity such as *playing a game* may appear in countless different ways in an image or video, depending on the kind of game, the environment, the players, and the type of recording. The Knowlywood knowledge base collects large amounts of activity knowledge and images from Hollywood movies, among other sources [37]. Conversely, statistics from large image and video collections can also improve natural language processing [33].

Another direction is to aim at more fine-grained knowledge, by not just classifying rectangular bounding boxes, but obtaining a detailed pixel-level analysis of shapes and contours. In fact, our ShapeLearner project [44] takes this one step further and provides pixel-level information about the parts of an object, e.g. distinguishing an animal's head from the rest of its body, or distinguishing the grip of a sword from its blade. ShapeLearner is thus both a knowledge graph and an image analysis engine.

User interfaces also greatly benefit from multimodality. Knowledge base entities can be visualized both temporally and geographically [10,11]. Queries may be multimodal as well. In a recent paper, we provided the first major steps towards multimodal question answering over Linked Data [14]. As mobile usage prevails, people now often have information needs that pertain to their surroundings and are best captured using an image.

4 Towards Cognitive and Neural Approaches

4.1 Neural Models

The ultimate goal of most knowledge bases is to enable more informed and intelligent applications. It has long been obvious that this will often require forms of inference that go beyond formal logical reasoning. For example, extracted and inferred knowledge assertions often come with confidence scores or probabilities that ought to be considered. In recent years, deep learning and other neural approaches have shown significant promise in this regard, enabling effective data-driven learning and inference for tasks that had just a few years go appeared intractable.

One important direction is to study semantic representations using neural methods. While well-known methods such as word2vec exploit the co-occurrence of words in large monolingual text corpora, recent work shows how to go beyond them and exploit further available cues in the data. One approach is to draw on information extraction to obtain higher-quality word embeddings [4]. We can also exploit document labels to learn high-quality representations for domain-specific concepts such as "carboplatin" or "prenatal exposure delayed effects" [16]. Additionally, it is now possible to obtain massively multilingual word representations covering many different languages simultaneously [7]. Last but not least, we can draw on large knowledge bases to learn embeddings for millions of entities in different languages [20].

Another direction is to investigate knowledge-driven applications of such representations in deep neural architectures. Currently, deep learning approaches are being investigated to discover salient information in text [45] and for neural information retrieval and ranking [12].

4.2 Common-Sense Knowledge

The final frontier is to go beyond learning towards genuinely intelligent behavior. This involves collecting substantial amounts of common-sense knowledge, which can take a number of different forms. We have investigated mining large amounts of basic commonsense knowledge assertions [39], fine-grained attributes [41], comparative commonsense knowledge [40] (e.g., that a falcon is faster than a leopard), and activity knowledge [37]. Such commonsense knowledge has been shown to aid in particularly challenging AI tasks such as metaphor interpretation [33].

However, human knowledge is unbounded and it is hence not sufficient to simply collect commonsense knowledge facts. For additional inference, we have investigated axiomatic rules [24] and large-scale reasoning [34,35]. Our latest approach is to combine commonsense knowledge with neural knowledge modeling, as exemplified by our WebBrain system [5]. WebBrain learns a neural model both from commonsense knowledge acquired from the Web as well as from general semantics as captured in word vector representations. With this knowledge, it attempts to make educated guesses beyond what has been observed on the Web. For example, WebBrain may guess that cockatiels are likely capable of flying, based on their similarity to other kinds of birds that fly.

5 Conclusion

While knowledge graphs have become widespread in industry and academia, we have seen that simple forms of structured facts do not fully resolve the traditional knowledge acquisition bottleneck.

In the future, systems will need to jointly learn and reason with knowledge from multiple heterogeneous sources of Big Data, including knowledge extracted from text, media, and large-scale structured knowledge repositories.

References

1. Bansal, M., Burkett, D., de Melo, G., Klein, D.: Structured learning for taxonomy induction with belief propagation. In: Proceedings of ACL 2014 (2014)
2. Basile, V., Cabrio, E., Schon, C.: KNEWS: Using logical and lexical semantics to extract knowledge from natural language. In: Proceedings of ECAI (2016)
3. Böhm, C., de Melo, G., Naumann, F., Weikum, G.: LINDA: distributed web-of-data-scale entity matching. In: Proceedings of CIKM 2012. ACM (2012)
4. Chen, J., Tandon, N., de Melo, G.: Neural word representations from large-scale commonsense knowledge. In: Proceedings of WI 2015 (2015)

5. Chen, J., Tandon, N., Hariman, C.D., de Melo, G.: WebBrain: joint neural learning of large-scale commonsense knowledge. In: Groth, P., Simperl, E., Gray, A., Sabou, M., Krötzsch, M., Lecue, F., Flöck, F., Gil, Y. (eds.) ISWC 2016. LNCS, vol. 9981, pp. 102–118. Springer, Cham (2016). doi:10.1007/978-3-319-46523-4_7
6. Corcoglioniti, F., Rospocher, M., Palmero Aprosio, A.: Frame-based ontology population with PIKES. TKDE **28**(12), 3261–3275 (2016)
7. de Melo, G.: Wiktionary-based word embeddings. In: Proceedings of MT Summit XV (2015)
8. Freitas, C., de Paiva, V., Rademaker, A., de Melo, G., Real, L., Silva, A.: Extending a lexicon of Portuguese nominalizations with data from corpora. In: Baptista, J., Mamede, N., Candeias, S., Paraboni, I., Pardo, T.A.S., Volpe Nunes, M.G. (eds.) PROPOR 2014. LNCS, vol. 8775, pp. 114–124. Springer, Cham (2014). doi:10.1007/978-3-319-09761-9_12
9. Gan, C., Lin, M., Yang, Y., de Melo, G., Hauptmann, A.G.: Concepts not alone: exploring pairwise relationships for zero-shot video activity recognition. In: Proceedings of AAAI. AAAI Press (2016)
10. Ge, T., Wang, Y., de Melo, G., Li, H.: Visualizing and curating knowledge graphs over time and space. In: Proceedings of ACL 2016. ACL (2016)
11. Hoffart, J., Suchanek, F.M., Berberich, K., Lewis-Kelham, E., de Melo, G., Weikum, G.: YAGO2: exploring and querying world knowledge in time, space, context, and many languages. In: Proceedings of WWW 2011. ACM (2011)
12. Hui, K., Yates, A., Berberich, K., de Melo, G.: A position-aware deep model for relevance matching in information retrieval. CoRR abs/1704.03940 (2017). http://arxiv.org/abs/1704.03940
13. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. Semant. Web **6**(2), 167–195 (2015)
14. Li, H., Wang, Y., de Melo, G., Tu, C., Chen, B.: Multimodal question answering over structured data with ambiguous entities. In: Proceedings of WWW 2017 (2017)
15. Long, X., Gan, C., de Melo, G.: Video captioning with multi-faceted attention. CoRR abs/1612.00234 (2016). http://arxiv.org/abs/1612.00234
16. Loza Mencía, E., de Melo, G., Nam, J.: Medical concept embeddings via labeled background corpora. In: Proceedings of LREC 2016 (2016)
17. McCrae, J.P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., de Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S., Osenova, P., Pareja-Lora, A., Pool, J.: The open linguistics working group: developing the linguistic linked open data cloud. In: Proceedings of LREC 2016 (2016)
18. de Melo, G.: Not quite the same: identity constraints for the Web of Linked Data. In: Proceedings of AAAI, pp. 1092–1098. AAAI Press (2013)
19. de Melo, G.: Lexvo.org: language-related information for the linguistic linked data cloud. Semantic Web **6**(4), 393–400 (2015)
20. de Melo, G.: Inducing conceptual embedding spaces from Wikipedia. In: Proceedings of WWW 2017. ACM (2017)
21. de Melo, G., Baker, C.F., Ide, N., Passonneau, R., Fellbaum, C.: Empirical comparisons of MASC word sense annotations. In: Proceedings of LREC 2012 (2012)
22. de Melo, G., Hose, K.: Searching the web of data. In: Serdyukov, P., Braslavski, P., Kuznetsov, S.O., Kamps, J., Rüger, S., Agichtein, E., Segalovich, I., Yilmaz, E. (eds.) ECIR 2013. LNCS, vol. 7814, pp. 869–873. Springer, Heidelberg (2013). doi:10.1007/978-3-642-36973-5_105

23. de Melo, G., de Paiva, V.: Sense-specific implicative commitments. In: Gelbukh, A. (ed.) CICLing 2014. LNCS, vol. 8403, pp. 391–402. Springer, Heidelberg (2014). doi:10.1007/978-3-642-54906-9_32
24. de Melo, G., Suchanek, F., Pease, A.: Integrating YAGO into the suggested upper merged ontology. In: Proceedings of ICTAI 2008. IEEE Computer Society (2008)
25. de Melo, G., Weikum, G.: Towards universal multilingual knowledge bases. In: Proceedings of the 5th Global WordNet Conference, pp. 149–156 (2010)
26. de Melo, G., Weikum, G.: Taxonomic data integration from multilingual Wikipedia editions. Knowl. Inf. Syst. **39**(1), 1–39 (2014)
27. de Paiva, V., Real, L., Rademaker, A., de Melo, G.: NomLex-PT: a lexicon of Portuguese nominalizations. In: Proceedings of LREC 2014. ELRA, May 2014
28. Rouces, J., de Melo, G., Hose, K.: FrameBase: representing N-Ary relations using semantic frames. In: Gandon, F., Sabou, M., Sack, H., d'Amato, C., Cudré-Mauroux, P., Zimmermann, A. (eds.) ESWC 2015. LNCS, vol. 9088, pp. 505–521. Springer, Cham (2015). doi:10.1007/978-3-319-18818-8_31
29. Rouces, J., de Melo, G., Hose, K.: Representing specialized events with FrameBase. In: Proceedings of the 4th International Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE) at ESWC 2015 (2015)
30. Rouces, J., de Melo, G., Hose, K.: Complex schema mapping and linking data: beyond binary predicates. In: Proceedings of LDOW 2016 (2016)
31. Rouces, J., de Melo, G., Hose, K.: Heuristics for connecting heterogeneous knowledge via FrameBase. In: Sack, H., Blomqvist, E., d'Aquin, M., Ghidini, C., Ponzetto, S.P., Lange, C. (eds.) ESWC 2016. LNCS, vol. 9678, pp. 20–35. Springer, Cham (2016). doi:10.1007/978-3-319-34129-3_2
32. Rouces, J., de Melo, G., Hose, K.: Klint: Assisting integration of heterogeneous knowledge. In: Proceedings of IJCAI 2016 (2016)
33. Shutova, E., Tandon, N., de Melo, G.: Perceptually grounded selectional preferences. Proceedings of ACL **2015**, 950–960 (2015)
34. Suda, M., Sutcliffe, G., Wischnewski, P., Lamotte-Schubert, M., de Melo, G.: External sources of axioms in automated theorem proving. In: Mertsching, B., Hund, M., Aziz, Z. (eds.) KI 2009. LNCS, vol. 5803, pp. 281–288. Springer, Heidelberg (2009). doi:10.1007/978-3-642-04617-9_36
35. Sutcliffe, G., Suda, M., Teyssandier, A., Dellis, N., de Melo, G.: Progress towards effective automated reasoning with world knowledge. In: Proceedings of the 23rd International FLAIRS Conference, pp. 110–115. AAAI Press (2010)
36. Tandon, N., de Melo, G.: Information extraction from web-scale n-gram data. In: SIGIR 2010 Web N-gram Workshop, vol. 5803, pp. 8–15. ACM (2010)
37. Tandon, N., de Melo, G., De, A., Weikum, G.: Knowlywood: mining activity knowledge from Hollywood narratives. In: Proceedings of CIKM 2015 (2015)
38. Tandon, N., de Melo, G., Suchanek, F.M., Weikum, G.: WebChild: harvesting and organizing commonsense knowledge from the web. In: Proceedings of WSDM. ACM (2014)
39. Tandon, N., de Melo, G., Weikum, G.: Deriving a Web-scale common sense fact database. In: Proceedings of AAAI, pp. 152–157. AAAI Press (2011)
40. Tandon, N., de Melo, G., Weikum, G.: Acquiring comparative commonsense knowledge from the web. In: Proceedings of AAAI, pp. 166–172. AAAI (2014)
41. Tandon, N., de Melo, G., Weikum, G.: WebChild 2.0: fine-grained commonsense knowledge distillation. In: Proceedings of ACL 2017. ACL (2017)
42. Wang, L., Cao, Z., de Melo, G., Liu, Z.: Relation classification via multi-level attention CNNs. In: Proceedings of ACL 2016 (2016)

43. Wang, Y., Ren, Z., Theobald, M., Dylla, M., de Melo, G.: Summary generation for temporal extractions. In: Hartmann, S., Ma, H. (eds.) DEXA 2016. LNCS, vol. 9827, pp. 370–386. Springer, Cham (2016). doi:10.1007/978-3-319-44403-1_23
44. Xu, H., Wang, Y., Feng, K., de Melo, G., Wu, W., Sharf, A., Chen, B.: Shape-learner: towards shape-based visual knowledge harvesting. In: Proceedings of ECAI 2016, pp. 435–443. IOS Press (2016)
45. Yang, Q., Cheng, Y., Wang, S., de Melo, G.: HiText: text reading with dynamic salience marking. In: Proceedings of WWW 2017. ACM (2017)
46. Yang, Q., Passonneau, R.J., de Melo, G.: PEAK: Pyramid evaluation via automated knowledge extraction. In: Proceedings of AAAI. AAAI Press (2016)