Con2KG - A Large-scale Domain-Specific Knowledge Graph

Nidhi Goyal
IIIT Delhi
New Delhi, India
nidhig@iiitd.ac.in

Niharika Sachdeva
Infoedge India Ltd.
Noida, India
niharika.sachdeva@naukri.com

Vijay Choudhary
Infoedge India Ltd.
Noida, India
vijay.choudhary@naukri.com

Rijula Kar
Infoedge India Ltd.
Noida, India
rijula.kar@naukri.com

Ponnurangam Kumaraguru
IIIT Delhi
New Delhi, India
pk@iiitd.ac.in

Nitendra Rajput
Infoedge India Ltd.
Noida, India
nitendra@naukri.com

ABSTRACT

This paper presents Con2KG, a large-scale recruitment domain Knowledge Graph that describes 4 million triples as facts from 250 thousands of unstructured data of job postings. We propose a novel framework for Knowledge Graph construction from unstructured text and an unsupervised, dynamically evolving ontology that helps Con2KG to capture hierarchical links between the entities missed by explicit relational facts in the triples. To enrich our graph, we include entity context and its polarity. Towards this end, we discuss Con2KG applications that may benefit the recruitment domain.

ACM Reference Format:
Nidhi Goyal, Niharika Sachdeva, Vijay Choudhary, Rijula Kar, Ponnurangam Kumaraguru, and Nitendra Rajput. 2019. Con2KG - A Large-scale Domain-Specific Knowledge Graph. In *30th ACM Conference on Hypertext & Social Media (HT '19), September 17–20, 2019, Hof, Germany*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3342220.3344931

1 INTRODUCTION

Knowledge Graph (KG) is a multi-relational structured graph which consists of various facts. Each fact includes entities represented as nodes and different relations represented as edges between them. KGs contribute tremendously in many applications such as search engines, AI assistants, etc. Online recruitment companies (such as LinkedIn, Indeed, etc.) also explore KG utility in various crucial tasks of job recruitment business such as personalized job suggestions, job search, candidate recommendation, content quality, etc. that can leverage the connected data. Despite the utility of KG in the recruitment domain, most of the existing Knowledge bases (KBs) such as Freebase [2], NELL [3], and DBpedia [6] provide limited facts which are of importance to recruitment domain. These KBs also lack essential entities such as evolving skills, designation, and hidden properties of the job, such as type of recruiter, shift timings, interview dates, etc. Prior Research works like T2KG [5] are specific to general concepts from day-to-day life and lack domain-specific knowledge. Given these challenges[1] and limitations, our work Con2KG proposes and demonstrates a framework for constructing large scale recruitment domain-specific KG of 4 million facts from 250 thousands of job postings. In contrast to the existing work, our contributions are:

- Con2KG exploits abundant information including properties such as skills, companies, work locations, type of job, type of company, shift timings, important dates, designation, candidate experience, type of qualification (degrees, diploma), and salary into a structure that helps recruiters and job seekers to organize knowledge about recruitment process.
- It provides a multi-tier architecture for the construction of KG from the structured and unstructured text from heterogeneous sources such as job postings, candidate profiles, etc.
- It offers a data-driven ontology mining of concepts and represents nuanced meanings of an entity when appearing within different contexts.

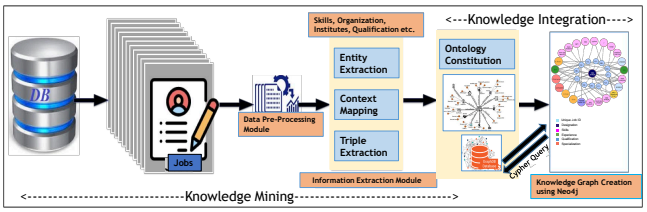

Figure 1: Architecture of Con2KG System

2 DATASET DESCRIPTION

We randomly sampled 250 thousand jobs from 1 million job postings collected over 120 months. We identified popular entities such as 12,057 skills, 60 qualifications, 87,905 institutes, 1,100 certifications, 2,23,955 companies, and 10,000 designations. We also extract essential facts like the type of job (home-based, full-time, etc.), recruiter type (company, consultancy, etc.), and shift timings (night, day, etc.) to form entities in our recruitment domain. We have 5,220 unique relations linking 3,65,061 entities extracted from unstructured and structured heterogeneous data.

[1] https://2018.semantics.cc/building-deploying-and-evolving-large-knowledge-graph-recruitment-domain-best-practices-and-lessons

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
HT '19, September 17–20, 2019, Hof, Germany
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6885-8/19/09.
https://doi.org/10.1145/3342220.3344931

3 SYSTEM OVERVIEW

Figure 1 gives an overview of the system architecture of Con2KG, which consists of two main modules:

3.1 Knowledge Mining Module

We divided this module into three phases:

- **Pre-Processing and Entity Extraction**: In this component, we first pre-process the noisy and unstructured data using NLP techniques. We employ sentence simplification module for complex sentences and part-of-speech approaches using Stanford Core NLP framework to revive missing phrases. Additionally, we use dependency parsing tree for rebuilding the syntactic and semantic structure [7]. To deal with abbreviations, we exploit rule-based heuristics and utilize a proprietary vocabulary list. Secondly, we discover the entities such as the type of company, recruiter type, important dates extraction, type of job using a hybrid combination of Stanford NER, libraries, dependency parser, and pattern-based heuristics.
- **Context mapping**: This component aims to describe the contextualization and polarities of the entities extracted from the previous steps. For Example, *"Candidate should be a Post-Graduate. Freshers cannot apply"*. We apply the entity extraction algorithm and remove 'Freshers'(negative polarity) from the list of entities extracted (Post-Graduate, Freshers). We also enhance our module by incorporating contextual information such as preferred candidates (Experienced, Fresher), etc. using Dependency Parser [7].
- **Triple Extraction and Ontology Constitution** : In this module, Firstly, we perform triple extraction using state-of-the-art OpenIE5 [8]. It extracts triples (subject, predicate, object) in a sentence by using pattern templates. In this, we identify the relations and its associated arguments in a sentence without using either prior domain knowledge. For Example, a sentence is *"Candidate should have experience of 4 years"*. After triple extraction we get *("Candidate", "should have experience of", "4 years")*. To add on missing entities, we also defined static relationships for the concepts whose triples are not extracted using OpenIE5. Secondly, we employ a hierarchical structure to these extracted key concepts and hidden nuances. We manually curated concepts related to companies, qualifications, etc. from the structured fields into the Ontology. We use the Louvain algorithm [1] to detect communities with the highest corresponding modularity and iteratively split these communities if the new partition had a positive modularity value. After ontology construction, we apply the state-of-the-art clustering approaches [4] to these entities capturing semantic information.

3.2 Knowledge Integration Module

In this module, we represent all of the extracted knowledge and store it into efficient graphical storage. All the entities (subjects and objects) are nodes and relationships are edges in the graph. Con2KG discovers and easily traverse through millions of nodes and edges using Neo4j Cypher Query Framework [10].

4 EVALUATION

We randomly selected 310 jobs from our legacy dataset containing 4719 sentences to evaluate the quality and quantity of the triples extracted. Based on the results, Con2KG can extract 1.72 triples per sentence on an average. We assess these triples and found 82% precision, 68.23% recall, and F-measure of 74.46%. We also analyze that triple extraction causes 0.05% errors due to incomplete triples and 0.20% due to no triple extraction for most of the sentences. Based on the preliminary analysis, errors in triple extraction occurs due to complexity in unstructured text and relations which are not identified clearly. Apart from these challenges, we still achieve approximately 74% both in terms of quantity and quality. Note that this evaluation methodology is also followed by T2KG [5].

5 FUTURE DIRECTIONS

Con2KG can exploit the entity and its relationships to predict and personalize the job suggestions in the recruitment domain using Entity Cards [9]. We can also query Con2KG and the complex data at the real-time to provide smart knowledge and detect false facts.

ACKNOWLEDGMENTS

This work is supported in part by SERB, FICII and InfoEdge India Limited. We would like to thank the members of the Analytics team (InfoEdge) and Precog, who gave us continued support throughout the project.

REFERENCES

[1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[2] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD '08)*. ACM, New York, NY, USA, 1247–1250. https://doi.org/10.1145/1376616.1376746

[3] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka, Jr., and Tom M. Mitchell. 2010. Toward an Architecture for Never-ending Language Learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI'10)*. AAAI Press, Palo Alto, California, USA, 1306–1313. http://dl.acm.org/citation.cfm?id=2898607.2898816

[4] Stephen C Johnson. 1967. Hierarchical clustering schemes. *Psychometrika* 32, 3 (1967), 241–254.

[5] Natthawut Kertkeidkachorn and Ryutaro Ichise. 2017. T2KG: An end-to-end system for creating knowledge graph from unstructured text. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI Press, Palo Alto, California, USA, 743–749.

[6] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 6, 2 (2015), 167–195.

[7] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, 55–60.

[8] Mausam Mausam. 2016. Open Information Extraction Systems and Downstream Applications. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, Palo Alto, California, USA, 4074–4077. http://dl.acm.org/citation.cfm?id=3061053.3061220

[9] Afroza Sultana, Quazi Mainul Hasan, Ashis Kumer Biswas, Soumyava Das, Habibur Rahman, Chris Ding, and Chengkai Li. 2012. Infobox Suggestion for Wikipedia Entities. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 2307–2310. https://doi.org/10.1145/2396761.2398627

[10] Jim Webber and Ian Robinson. 2018. *A Programmatic Introduction to Neo4J* (1st ed.). Addison-Wesley Professional, Boston, USA.