LANGUAGE, STRUCTURE, TIME AWARE KNOWLEDGE BASE COMPLETION

PRACHI JAIN

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
October 2022

LANGUAGE, STRUCTURE, TIME AWARE KNOWLEDGE BASE COMPLETION

by

PRACHI JAIN

Department of Computer Science and Engineering

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

October 2022

This dissertation is dedicated to
my loving grandparents
and
my son Arinjay – may you never stop learning.

Certificate

This is to certify that the dissertation titled **Language, Structure, Time aware Knowledge Base Completion** being submitted by **Ms Prachi Jain** for the award of **Doctor of Philosophy** in Department of Computer Science and Engineering is a record of bonafide work carried out by her under my guidance and supervision at the **Department of Computer Science and Engineering, Indian Institute of Technology Delhi**. The work presented in this dissertation has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma unless otherwise stated explicitly. In particular, work done in Chapters 4, 5, 6 and 7 were done jointly with undergraduate students. In each case, the part done by the collaborators appeared in their respective bachelor's thesis.

Mausam
Professor
Department of Computer Science and Engg.
Indian Institute of Technology Delhi
New Delhi- 110016

Soumen Chakrabarti
Professor
Department of Computer Science and Engg.
Indian Institute of Technology Bombay
Powai, Mumbai, Maharashtra- 400076

Acknowledgements

*"***It's all about the journey, not the outcome.***"*

Carl Lewis

As I am about to end my past-due PhD journey. I want to look back and extend my gratitude to all the people whom I met in the course of my PhD, without whom the long journey would have looked very dull.

I would like to express my deepest gratitude to my advisors Mausam and Soumen for their guidance and support during my PhD. Mausam's drive to do high-quality research and a constant push to go beyond the limits always kept me on my toes. His honest critical feedback always empowered me to be the best I can be. On the other hand, Soumen's passion for the subject and ability to pay attention to the minutest of details always kept me motivated and pushed me to do my best. I have vastly benefited from this unique mix of advisors. In these last two years, there were times when work life was unavoidably affected by events outside work. I am so grateful to Mausam and Soumen for caring, understanding, and prioritizing their students' well-being. Thank you Mausam and Soumen, for patiently being there in this period of my life and my career and always showering me with the wisest of advice.

I am thankful to Dr Denial Bernhardt and Apoorv Saxena for hosting me for an internship at Facebook. I would also like to thank Dr Sebastian Riedel, Sara Khodeir, and Gerard Goossen for their constant feedback and support during the internship.

I also want to acknowledge the TCS PhD fellowship for supporting me financially during my PhD. I am also thankful to Microsoft Research, Xerox Research, Google Research, IJCAI, and ACL student travel grants for supporting my travel to numerous conferences at various points in my PhD. I would also like to acknowledge the administrative staff (especially Rekha mam and Hemant bhaiya) at the Computer Science Department in IIT Delhi for their continuous help throughout these years. I was very lucky to have collaborated with some of the brightest undergraduate students on campus — Shikhar Murty, Kabir Chhabra, Pankaj Kumar, Mukund Mundhra, Sushant Rathi. I learnt a lot from them.

I would like to thank the organizing committee of the CIFAR Deep Learning + Reinforcement Learning (DLRL) Summer School and the Lisbon School of Machine Learning Summer

School (LxMLS) for organizing amazing summer schools. I learned a lot and made good friends during the summer schools. I would also like to thank them for supporting my travel. I will always remember the interesting discussions and fun trips with Disha during my time at DLRL summer school.

I have made lifelong friends over the last few years – Ankit, Happy, Himanshu and Yashoteja (aka JHAMP). Thank you for all the laughter, food, and wisdom to keep me moving over these years. I will always look upto them for any professional or personal advise. There are many more wonderful friends I made in my PhD — Yatin, Keshav, Vishal, Dilpreet, Jatin, Chandrika, Deepak, Madhulika, Prajna and Sanchi.

Finally, I would like to express my heartfelt gratitude to my parents for the numerous sacrifices they have made to support me and my education. Their unwavering faith always helped me sail through all the ups and downs of my PhD journey. I would also like to thank my grandparents, husband, sister, brother and parent-in-laws, without their love, support and encouragement I would not have achieved this feat.

Abstract

Large Knowledge Bases (KBs) have been built to access a comprehensive collection of facts in a machine readable automatic format. Although these KBs are large, their coverage is far from complete. Most relations between entities are found to be missing in many widely-used KBs. Inference can be used to improve the coverage of such KBs and hence make them more suitable for practical applications like search, dialogue and question answering. This inference process is called Knowledge Base Completion (KBC).

In this dissertation, we analyze existing KBC systems and propose various new KBC methods and models. In particular, we exploit various attributes of KBs — language, structure and temporal attributes to improve KBC performance. We also extensively study the evaluation of KBC models and propose fair evaluation policies.

First, we studied different aspects of modeling KB structure and their impact on KBC performance:

- KBC models can be categorized on the basis of the way they represent entities. Matrix factorization (MF) models have a vector defined for each entity-*pair*, while tensor factorization (TF) models maintain a vector for each *entity*. We compared the effectiveness of the MF and TF paradigms for the general task of KBC. We recognized that special care is needed to handle out-of-vocabulary entity-pairs when evaluating MF against TF. We also propose the first fair unified KBC evaluation protocol to compare MF and TF approaches for KBC.

- Our analysis of KBC models reveals that they often make entity predictions that are incompatible with the type required by the relation. For example, DistMult incorrectly predicts *'Akira Isida'* (type-person) for the query *'Chief Phillips (type-film), released_in_region, ?'*. We propose an unsupervised typing gadget, which enhances KBC models (like DistMult and Complex) with type-compatibility checkers. The enhanced models (TypeDM and TypeComplEx) showed improved KBC performance over the base models. Further analysis revealed that our models better represent the latent types of entities and their embeddings also predict supervised types better than the embeddings learned by baseline models.

While implementing the above models, we found that the norm (L1, L2, or L3) used for regularization or measuring distances, and the rate of negative sampling used to train the models, can have significant consequences for KBC accuracy, sometimes overturning conventional wisdom about how various models compare with each other. Through our investigation of these issues, we implemented a very competitive version of ComplEx KB embedding, better than some follow-up systems.

Next, we study temporal KBs, which associate a relational fact (s, r, o) with a valid set of times (often an instant or interval). We propose TIMEPLEX a Temporal Knowledge Base Completion (TKBC) model, primarily targeted to the link-prediction and time-interval prediction tasks. To the best of our knowledge, this is the first work that predicts the time interval in which the given fact is valid in a general model-independent manner. Also, this is the first work that proposes a time-aware evaluation strategy for TKBC.

Finally, we study Open KBs where entities and relations are represented via textual schema-free strings. Open KBC is generally performed using an inference rule corpus. Using linguistic insights, we develop an algorithm —- Knowledge Guided Linguistic Rewrites (KGLR) — which provides independent verification for statistically-generated Open KB inference rules. The generated high precision rule corpus eventually helps in improving the KBC task performance.

सार

तथ्यों के व्यापक संग्रह को एक मशीन पठनीय स्वचालित प्रारूप मैं लाने के लिए बड़े ज्ञानकोष (नॉलेज बेस - के.बी.) का निर्माण किया गया है। हालांकि ये के.बी. बड़े हैं, लेकिन इनका कवरेज अधूरा है। कई व्यापक रूप से उपयोग किए जाने वाले के.बी. में इकाइयों (एंटिटी) के बीच अधिकांश संबंध (रिलेशन) गायब पाए जाते हैं। ऐसे के.बी. के कवरेज में सुधार के लिए अनुमान (इनफ्रेंस) का उपयोग किया जा सकता है और इससे उन्हें व्यावहारिक अनुप्रयोगों (जैसे खोज, संवाद और प्रश्न उत्तर) के लिए अधिक उपयुक्त बनाया जा सकता है। यह इनफ्रेंस प्रक्रिया को नॉलेज बेस कंप्लीशन (के.बी.सी.) कहा जाता है।

इस थीसिस में, हम मौजूदा के.बी.सी. सिस्टम का विश्लेषण करते हैं और विभिन्न नई के.बी.सी. विधियों का प्रस्ताव करते हैं। विशेष रूप से, हम के.बी.सी. प्रदर्शन में सुधार करने के लिए के.बी. की विभिन्न विशेषताओं का फायदा उठाते हैं - भाषा, संरचना और समय। हम के.बी.सी. मॉडल के मूल्यांकन का भी बड़े पैमाने पर अध्ययन करते हैं और निष्पक्ष मूल्यांकन नीतियों का प्रस्ताव रखते है।

सबसे पहले, हमने के.बी. संरचना के मॉडलिंग के विभिन्न पहलुओं और के.बी.सी.पर उनके प्रभाव का अध्ययन किया।

- के.बी.सी. मॉडल को उनके द्वारा एंटिटीज का प्रतिनिधित्व करने के तरीके के आधार पर वर्गीकृत किया जा सकता है। मैट्रिक्स फ़ैक्टराइज़ेशन (एम.एफ.) मॉडल में प्रत्येक इकाई-जोड़ी (एंटिटी पेयर) के लिए एक वेक्टर परिभाषित होता है, जबकि टेंसर फ़ैक्टराइज़ेशन (टी.एफ.) मॉडल प्रत्येक इकाई (एंटिटी) के लिए एक वेक्टर बनाए रखते हैं। हमने एम.एफ. और टी.एफ. की प्रभावशीलता की तुलना के.बी.सी. में प्रदर्शन से की। हमने माना कि एम.एफ. और टी.एफ. की प्रभावशीलता की तुलना करते समय आउट-ऑफ-वोकैबुलरी इकाई-जोड़े को विशेष देखभाल की आवश्यकता होती है। और हम एम.एफ. और टी.एफ. दृष्टिकोणों की तुलना करने के लिए पहले निष्पक्ष एकीकृत के.बी.सी. मूल्यांकन प्रोटोकॉल का प्रस्ताव करते है।
- के.बी.सी. मॉडल के हमारे विश्लेषण से पता चलता है कि वे अक्सर ऐसी भविष्यवाणी करते हैं जो असंगत होती हैं संबंध द्वारा आवश्यक प्रकार के साथ। उदाहरण के लिए, DistMult गलत भविष्यवाणी करता है - ''अकीरा इसिदा (टाइप-पर्सन)" की, प्रश्न - ''चीफ फिलिप्स (टाइप-फिल्म)", "क्षेत्र में रिलीज", ?' के पूछे जाने पर।

- हम एक गैर-पर्यवेक्षित टाइपिंग गैजेट का प्रस्ताव करते हैं, जो के.बी.सी. मॉडल (जैसे DistMult और ComplEx) को बेहतर करता है टाइप-संगतता चेकर्स के साथ। एन्हांस्ड मॉडल (TypeDM और TypeComplEx) ने बेस मॉडल की तुलना में बेहतर के.बी.सी. प्रदर्शन दिखाया। आगे विश्लेषण से पता चला कि हमारे मॉडल अव्यक्त प्रकार की एंटिटीज का बेहतर प्रतिनिधित्व करते हैं और उनके एम्बेडिंग भी पर्यवेक्षित प्रकारों की भविष्यवाणी बेसलाइन मॉडल द्वारा सीखी गई एम्बेडिंग से बेहतर है ।

उपरोक्त मॉडलों को लागू करते समय, हमने पाया कि नॉर्म के लिए प्रयुक्त मानदंड (L1, L2, या L3) या दूरियों को मापने, और मॉडलों को प्रशिक्षित करने के लिए उपयोग किए जाने वाले नकारात्मक नमूने की दर, केबीसी सटीकता के लिए महत्वपूर्ण परिणाम हो सकते हैं, यह कभी-कभी विभिन्न मॉडलों की एक दूसरे के साथ तुलना के पारंपरिक ज्ञान को उलट देते हैं। इनकी जांच के माध्यम से, हमने ComplEx के.बी. एम्बेडिंग का एक बहुत ही प्रतिस्पर्धी संस्करण लागू किया है।

इसके बाद, हम टेंपोरल के.बी. का अध्ययन करते हैं, जो एक तथ्य (s, r, o) को एक वैध समय के सेट (अक्सर एक पल या अंतराल) के साथ जोड़ते हैं। हम टी.के.बी.सी. के लिए TimePlex मॉडल प्रस्तावित करते है, यह मॉडल मुख्य रूप से लिंक-भविष्यवाणी और समय-अंतराल भविष्यवाणी के लिए लक्षित है। हमारी सर्वोत्तम जानकारी के अनुसार, यह पहला काम है जो समय अंतराल की भविष्यवाणी करता है जिसमें दिया गया तथ्य सामान्य मॉडल-स्वतंत्र तरीके से मान्य है। साथ ही, यह पहला काम है जो टी.के.बी.सी.के लिए समय के प्रति जागरूक मूल्यांकन रणनीति का प्रस्ताव करता है।

अंत में, हम ओपन के.बी. का अध्ययन करते हैं जहां स्कीमा-मुक्त टेक्स्ट के माध्यम से एंटिटीज और रिलेशंस का प्रतिनिधित्व किया जाता है। ओपन के.बी.सी. आमतौर पर एक अनुमान (इनफ्रेंस) नियम कोष का उपयोग करके किया जाता है।

हम एक एल्गोरिथ्म विकसित करते हैं – ज्ञान निर्देशित भाषाई पुनर्लेखन (KGLR) – जो सांख्यिकीय रूप से उत्पन्न ओपन केबी अनुमान नियमों के लिए स्वतंत्र सत्यापन प्रदान करता है। भाषा की अंतर्दृष्टि के प्रयोग से हम एक एल्गोरिथम विकसित करते हैं - ज्ञान निर्देशित भाषाई पुनर्लेखन (के.जी.एल.आर.) -जो सांख्यिकीय रूप से उत्पन्न ओपन के.बी. अनुमान (इनफ्रेंस) नियमों के लिए स्वतंत्र सत्यापन प्रदान करता है। उत्पन्न उच्च परिशुद्धता नियम कॉर्पस अंततः के.बी.सी. कार्य को बेहतर बनाने में मदद करता है।

Contents

List of Figures

List of Tables