NEURAL METHODS FOR MONOLINGUAL AND MULTILINGUAL OPEN INFORMATION EXTRACTION

KESHAV SAI KOLLURU

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
JULY 2023

©Indian Institute of Technology Delhi - 2023
All rights reserved.

NEURAL METHODS FOR MONOLINGUAL AND MULTILINGUAL OPEN INFORMATION EXTRACTION

by

KESHAV SAI KOLLURU

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
JULY 2023

Certificate

This is to certify that the thesis titled **Neural Methods for Monolingual and Multilingual Open Information Extraction** being submitted by **Mr. Keshav Sai Kolluru** for the award of **Doctor of Philosophy** in Department of Computer Science and Engineering is a record of bonafide work carried out by him under my guidance and supervision at the **Department of Computer Science and Engineering, Indian Institute of Technology Delhi**. Unless otherwise stated explicitly, the work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma. In particular, some parts of work done in Chapters 3 and 6 was done jointly with undergraduate students and some part of Chapter 7 was done jointly with a master's student. In each case, the part done by the collaborators appeared in their respective theses.

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
Mumbai- 400076

Acknowledgements

I am profoundly grateful to my advisors, Prof Mausam and Prof Soumen Chakrabarti, whose knowledge, wisdom, and mentorship have guided me unfailingly throughout the challenging but rewarding journey that is a PhD. Their keen insights, tireless dedication, and unwavering belief in my capabilities have played a pivotal role in my academic growth and accomplishment. I am grateful for their patience, constant encouragement, and invaluable advice.

Undertaking a PhD is a long, tough journey, akin to sailing in the vast ocean where one often encounters unanticipated storms and calm, peaceful stretches alike. It tests your strength, resilience, and determination in ways you never anticipated, yet you emerge stronger, wiser, and more resilient at the end of this voyage. You arrive at the shore equipped with a trove of knowledge, skills, and experiences that fortify your capabilities for future endeavours. In this profound journey, the strength I gained and the person I have become, I owe to many who have helped me along the way. I am particularly grateful to my family, who have always stood by me through thick and thin; all my teachers, from whom I have learnt invaluable lessons; my friends who helped me become a better person; and colleagues who went out of their way to help me.

Keshav Kolluru
July 2023

Abstract

Open Information Extraction (Open IE) aims to extract semi-structured information from natural language text in a domain-independent fashion. It is formulated as extracting a set of tuples of the form (subject, relation, object) where each of the fields corresponds to a phrase in the text. Compared to 'closed' information extraction based on canonical KGs, it avoids the need for experts to define the ontology and data curators, making it scalable across domains. In this dissertation, we describe novel Open IE systems that take advantage of recent advances in deep neural models to tackle multiple challenges associated with building automated systems for the task of Open IE in both monolingual and multilingual settings. We propose solutions that represent significant advances across multiple axes — (1) design of new models, (2) extension to multiple languages, (3) support for linguistic phenomena, (4) downstream application to Knowledge Bases and (5) release of new systems. In models, we build novel deep learning architectures that establish new state-of-art performance by faithful modelling of the Open IE task with pre-trained language models. We experiment with both sequence-to-sequence generation models (named IMoJIE, Gen2OIE) and sequence labeling models (named IGL, CIGL) for the task. IMoJIE (Iterative Memory-based Joint Open Information Extraction) iteratively re-encodes the sentence along with the extractions generated so far to generate the remaining extractions, ensuring diversity in the extractions. Gen2OIE is a two-stage generative model that first generates all the relations in the sentence, followed by generating extractions corresponding to each relation. The IGL (Iterative Grid Labeling) model labels all the words in the sentence in an iterative fashion with tags dictating their position in the Open IE tuples. CIGL improves over IGL by adding constraints in training to increase the coverage of the extractions. In multilinguality, to enable extension of Open IE to other languages, we need training data in the respective language. Therefore, we build a pipeline for translating English Open IE training data and generating high-quality data in Spanish, Portuguese, Chinese, Hindi and Telugu. In linguistic phenomena, noticing that current Open IE systems lack in properly handling certain linguistic phenomena such as noun compounds and conjunctions, we develop systems for noun compound interpretation and coordination analysis which are incorporated into Open IE systems. In applications of Open IE extractions, we build a multilingual fact linking benchmark and model for connecting textual extractions to their knowledge bases while accounting for facts that can exist in multiple languages. In another application, we advance the state of art in Open Knowledge Base Completion by using a two-stage entity-aware pipeline to infer new triples. Finally in systems, we release the OpenIE-6 system that represents the cutting-edge in the line of Open IE software packages.

सार

खुली सूचना निष्कर्षण (ओपन आईई) का उद्देश्य प्राकृतिक भाषा में लिखा हुआ पाठ से क्षेत्र-स्वतंत्र रूप में अर्ध-संरचित जानकारी निकालना है। ओपन आईई (विषय, संबंध, वस्तु) के सेट निकालने के रूप में परिभाषित किया गया है, जहां प्रत्येक फ़ील्ड पाठ में एक वाक्यांश से मेल खाती है। कैनोनिकल केजी पर आधारित 'बंद' सूचना निष्कर्षण की तुलना में, यह ऑन्तोलॉजी और डेटा क्यूरेटर्स की आवश्यकता से बचाता है, जिससे यह सभी क्षेत्र में काम कर लेता है। इस शोध प्रबंध में, हम नए ओपन आईई सिस्टम्स का वर्णन करते हैं जो एकभाषी और बहुभाषी दोनों ही सेटिंग में ओपन आईई के कार्य के लिए ऑटोमेटेड सिस्टम बनाता है। हम ऐसे समाधान प्रस्तावित करते हैं जो कई अक्षों में महत्वपूर्ण प्रगति का प्रतिनिधित्व करते हैं --- (1) नए मॉडलों का डिज़ाइन, (2) कई भाषाओं में विस्तार, (3) भाषाई घटनाओं के लिए समर्थन, (4) नॉलेज बेस के डाउनस्ट्रीम अनुप्रयोग और (5) नए सिस्टम की रिलीज। मॉडलों में, हम नए डीप लर्निंग आर्किटेक्चर बनाते हैं जो प्री-ट्रेनिंग लैंग्वेज मॉडल के साथ ओपन आईई टास्क का ईमानदारी से मॉडलिंग करके नए अत्याधुनिक प्रदर्शन स्थापित करते हैं। हम इस काम के लिए सीक्वेंस-टू-सीक्वेंस जेनरेशन मॉडल (आईएमओजेई, जेन2ओईई नाम) और सीक्वेंस लेबलिंग मॉडल (आईजीएल, सीआईजीएल नाम) दोनों के साथ प्रयोग करते हैं। इमोजी (इटरेटिव मेमोरी-आधारित ज्वाइंट ओपन इंफॉर्मेशन एक्सट्रैक्शन) शेष निष्कर्षण उत्पन्न करने के लिए अब तक उत्पन्न निष्कर्षण के साथ वाक्य को जोड़ता है, जिससे निष्कर्षण में विविधता सुनिश्चित होती है। जेन2ओईई एक दो-स्टेज वाला जेनरेटिव मॉडल है जो पहले वाक्य में सभी संबंध उत्पन्न करता है, उसके बाद हर संबंध के अनुरूप निष्कर्षण उत्पन्न करता है। आईजीएल (इटरेटिव ग्रिड लेबलिंग) मॉडल वाक्य में सभी शब्दों को ओपन आईई टैग के साथ पुनरावृत्त फैशन में लेबल करता है। सीआईजीएल निष्कर्षण का कवरेज बढ़ाने के लिए ट्रेनिंग में प्रतिबंध लगाकर आईजीएल पर सुधार करता है। ओपन आईई के निष्कर्षण को अन्य भाषाओं में सक्षम करने के लिए, हमें संबंधित भाषा में ट्रेनिंग डेटा चाहिए. इसलिए, इंग्लिश ओपन आईई ट्रेनिंग डेटा का अनुवाद करके स्पेनिश, पुर्तगाली, चीनी, हिंदी और तेलुगु में उच्च क्वालिटी का डेटा उत्पन्न करने के लिए हम पाइपलाइन बनाते हैं। भाषाई घटनाओं में, इस बात पर ध्यान देते हुए कि मौजूदा ओपन आईई सिस्टम कुछ भाषाई घटनाओं जैसे कि नोन कंपाउंड और कंजंक्शन को सही तरीके से संभालने में कमी है, हम नोन कंपाउंड व्याख्या और समन्वय विश्लेषण के लिए सिस्टम विकसित करते हैं जो ओपन आईई सिस्टम में शामिल किया जाता है। ओपन आईई निष्कर्षण के अनुप्रयोगों में, हम एक बहुभाषी तथ्य लिंकिंग बेंचमार्क और मॉडल बनाते हैं, जो कई भाषाओं में मौजूद तथ्यों का हिसाब करते हुए उनके ज्ञान के आधार पर नॉलेज बेस और निष्कर्षण को जोड़ सकता है। एक अन्य अनुप्रयोग में, हम नए निष्कर्षण का अनुमान लगाने के लिए पाइपलाइन का उपयोग करके ओपन नॉलेज बेस कंपलेशेशन में अत्याधुनिक तकनीक को आगे बढ़ाते हैं। अंततः सिस्टम में, हम 'ओपनआईई-६' सिस्टम जारी करते हैं जो 'ओपन आईई' सॉफ्टवेयर पैकेजों की श्रेणी में अत्याधुनिक का प्रतिनिधित्व करता है।

Contents

1 Introduction **1**
- 1.1 Semi-structured nature of Open IE . . . 2
- 1.2 Relevance of Open IE . . . 3
- 1.3 Thesis Contributions . . . 4
  - 1.3.1 Models . . . 5
    - 1.3.1.1 Generation models . . . 6
    - 1.3.1.2 Labeling models . . . 6
  - 1.3.2 Linguistic Phenomena . . . 7
    - 1.3.2.1 Conjunctions . . . 7
    - 1.3.2.2 Proper Noun Compounds . . . 7
  - 1.3.3 Multilinguality . . . 8
  - 1.3.4 Applications . . . 8
  - 1.3.5 Systems . . . 9
- 1.4 Thesis Outline . . . 9

2 Related Work **10**
- 2.1 Task Definition . . . 10
- 2.2 Evaluation . . . 11
- 2.3 Models for English Open IE . . . 12
  - 2.3.1 Syntactic and Statistical Models . . . 12
    - 2.3.1.1 TextRunner . . . 13
    - 2.3.1.2 ReVerb . . . 13
    - 2.3.1.3 OLLIE . . . 13
    - 2.3.1.4 StanfordIE . . . 14
    - 2.3.1.5 ClausIE . . . 14
    - 2.3.1.6 OpenIE-4 . . . 14
    - 2.3.1.7 OpenIE-5 . . . 15
    - 2.3.1.8 MinIE . . . 15
    - 2.3.1.9 NestIE . . . 16
  - 2.3.2 Deep Learning Models . . . 16
    - 2.3.2.1 RnnOIE . . . 17
    - 2.3.2.2 SenseOIE . . . 17
    - 2.3.2.3 Iterative Rank-Aware Learning . . . 17
    - 2.3.2.4 SpanOIE . . . 18
    - 2.3.2.5 Systematic Comparison . . . 18
    - 2.3.2.6 CopyAttention . . . 19
    - 2.3.2.7 MCTS . . . 19
    - 2.3.2.8 DocOIE . . . 19

2.4 Models for Non-English Open IE . . . 20
2.4.1 Open IE models for German . . . 20
2.4.2 Open IE models for Italian . . . 20
2.4.3 Open IE models for Greek . . . 21
2.4.4 Open IE models for Chinese . . . 21
2.4.4.1 Logician . . . 21
2.4.4.2 Orator . . . 22
2.4.5 Models for multilingual Open IE . . . 22
2.4.5.1 Cross Lingual Projection (CLP) . . . 22
2.4.5.2 PredPatt . . . 24
2.4.5.3 ArgOIE . . . 24
2.4.5.4 CrossOIE . . . 24
2.4.5.5 Multi2OIE . . . 24
2.5 Applications of Open IE . . . 25
2.5.1 Text Summarization . . . 25
2.5.2 Question Answering . . . 25
2.5.3 Event Extraction . . . 25
2.5.4 Entity and Relation Linking . . . 26
2.5.5 Video Grounding . . . 26
2.5.6 Scientific Text . . . 26
2.6 Related Tasks . . . 27
2.6.1 Ontological/Closed IE . . . 27
2.6.2 Semantic Role Labeling . . . 28
2.6.3 Open Link Prediction . . . 28
2.6.4 Canonicalization . . . 28

3 Generative Models for Open IE **30**
3.1 IMoJIE: Iterative Memory Joint Open Information Extraction . . . 30
3.1.1 Confidence Scoring . . . 33
3.2 Gen2OIE: Two-Stage Generative Model . . . 33
3.2.1 Confidence Scoring . . . 35
3.3 Experimental Setup . . . 35
3.3.1 Training Data Construction . . . 35
3.3.2 Evaluation Metric . . . 35
3.3.3 Systems Compared . . . 36
3.3.4 Implementation . . . 37
3.4 Results and Analysis . . . 37
3.4.1 Performance of IMoJIE . . . 37
3.4.2 Performance of Gen2OIE . . . 38
3.4.3 Redundancy . . . 38
3.4.4 Performance with varying sentence lengths . . . 40
3.4.5 Effectiveness of pre-trained decoders . . . 40
3.4.6 Discussion on Order of Extractions . . . 41
3.5 Conclusion . . . 41

4 Labeling Models for Open IE **43**
4.1 Iterative Grid Labeling for Open IE . . . 44
4.2 Grid Constraints . . . 46
4.2.1 POS Coverage (**POSC**) . . . 46
4.2.2 Head Verb Coverage (**HVC**) . . . 46
4.2.3 Head Verb Exclusivity (**HVE**) . . . 47
4.2.4 Extraction Count (**EC**) . . . 47
4.3 Confidence Rescoring . . . 47
4.4 Experimental Setup . . . 48
4.5 Experiments . . . 48
4.5.1 Speed and Performance . . . 49
4.5.2 Constraints Ablation . . . 49
4.5.3 Performance using different metrics . . . 52
4.6 Conclusion . . . 52

5 Handling of Linguistic Phenomena in Open IE **53**
5.1 Coordinations . . . 53
5.1.1 Coordination Analyzer . . . 54
5.1.1.1 Experimental Setup . . . 55
5.1.1.2 Experiments . . . 55
5.1.2 Coordination Analyzer in Open IE . . . 55
5.1.2.1 Evaluation . . . 56
5.1.2.2 Experiments . . . 57
5.1.2.3 Manual Comparison . . . 58
5.1.3 Discussion . . . 59
5.2 Proper Noun Compound Interpretation . . . 60
5.2.1 Related Work . . . 61
5.2.2 Problem Definition . . . 62
5.2.3 PRONCI Dataset . . . 62
5.2.4 Models . . . 64
5.2.5 Experimental Setup . . . 65
5.2.6 Experimental Results . . . 67
5.2.6.1 Performance of Supervised Models . . . 67
5.2.6.2 Performance of few-shot learning . . . 69
5.2.6.3 Proper noun vs. Common noun . . . 70
5.2.6.4 Quality Assessment of Evaluation Metrics . . . 70
5.2.6.5 Random Split of PRONCI . . . 71
5.2.6.6 Effect of Pretraining . . . 72
5.2.6.7 Adding multiple sources of knowledge . . . 72
5.2.6.8 Error Analysis . . . 73
5.2.7 Application to Open IE . . . 73
5.3 Open IE Systems: Open IE 6.2 . . . 74

6 Interlingual Transfer of Open IE Training Data **76**
6.1 Alignment Augmented Consistent Translation . . . 76
6.2 AACTrans: Crosslingual Data Transfer . . . 77
6.2.1 Consistent Translation . . . 78
6.2.2 Consistent Translation for Crosslingual Data Transfer . . . 79

6.2.3 Crosslingual Label Projection (CLP) . . . 80
6.3 Experimental Setting . . . 80
6.4 Experiments . . . 81
6.4.1 Quality of AACTrans+CLP data . . . 82
6.4.2 Evaluating Consistency . . . 83
6.4.3 Ablation Study . . . 84
6.4.4 BLEU scores . . . 84
6.4.5 Effect of word alignments quality . . . 85
6.5 Conclusion . . . 85

7 Application of Open IE to Knowledge Bases 86
7.1 Knowledge Base Fact Linking . . . 86
7.1.1 Related Work . . . 89
7.1.2 Multilingual Fact Linking: Problem Overview . . . 89
7.1.3 IndicLink: A New Dataset for Fact Linking in Indian Languages . . . 90
7.1.4 ReFCoG: Proposed Method for MFL . . . 90
7.1.4.1 Fact-Text Dual Encoder for Retrieval . . . 91
7.1.4.2 Cross Encoders for Re-ranking . . . 91
7.1.5 Experimental Setting . . . 93
7.1.6 Experiments . . . 93
7.1.6.1 Effectiveness of ReFCoG . . . 94
7.1.6.2 ReFCoG ablations . . . 94
7.1.6.3 Effect of Multilingual Fact Surface Forms . . . 95
7.1.6.4 ReFCoG Error Analysis . . . 95
7.1.7 Effectiveness of ReFCoG for linking Open IE tuples . . . 96
7.2 Open Knowledge Base Completion . . . 97
7.2.1 Related Work . . . 98
7.2.2 CEAR: Cross-Entity Aware Reranker . . . 98
7.2.3 Experimental Setting . . . 100
7.2.4 Experiments . . . 101
7.3 Conclusion . . . 101

8 Conclusion and Future Work 103
8.1 Non-Autoregressive models . . . 104
8.2 Large-scale multilingual support . . . 104
8.3 Evaluation metrics . . . 104
8.4 Downstream Applications . . . 105
8.5 Implicit Relations . . . 105
8.6 Entity and Relation Canonicalization . . . 106
8.7 User-Facing tasks . . . 106
8.8 Customizability . . . 106

Biography 120

List of Figures

1.1 Open Information Extraction (Open IE) systems extract tuples of the format (subject, relation, object) from a sentence. A collection of such tuples form an Open Knowledge Base, which can be used as a source of factual information. They provide additional value over using raw-text due to the possibility of aggregating extractions from multiple source sentences via clustering (Fan et al., 2019a). 2

1.2 Schematic of the overall contributions of the thesis. We introduce new Open IE models, which are generation-based (IMoJIE, Gen2OIE) in Chapter 3 and labeling-based (IGL, CIGL) in Chapter 4.We handle special linguistic phenomena in Open IE extractions such as noun compounds (NCI) coordination analysis (Coord-IGL) in Chapter 5. We extend Open IE to other languages by creating a novel training data translation technique (AACTrans) in Chapter 6. We use Open IE in downstream applications of multilingual fact linking (MFL) and Open KB completion (CEAR) in Chapter 7. 5

2.1 Equivalent English and Spanish sentence with corresponding word alignments between them 23

2.2 Equivalent English and Spanish sentence with corresponding word alignments between them 23

3.1 One step of the sequential decoding process, for generating the i^{th} extraction, which takes the original sentence and all extractions numbered $1, \ldots, i-1$, previously generated, as input. 32

3.2 Gen2OIE model contains two Seq2Seq models. In Stage-1, it generates all relations in the sentence, separated by an [SEP] token. For each detected relation in Stage-2, it generates extractions containing the relation. 34

3.3 Precision-Recall curve of Open IE Systems. 38

3.4 Measuring performance with varying input sentence lengths 40

4.1 The extractions *(Rome; [is] the capital of; Italy)* and *(Rome; is known for; it's rich history)* can be seen as the output of grid labeling. We additionally introduce a synthetic token *[is]* to the input to facilitate more natural relation extractions. 43

4.2 2-D grid for Open IE with extraction as rows and words as columns. The values represent the labels (*S*)ubject, (*R*)elation, (*O*)bject. The empty cells represent *None*. Constraints can be applied across rows and columns. 44

4.3 Architecture of IGL. BERT-embeddings of the words are iteratively passed through self-attention layers. st_1, st_2, st_3 refer to the appended tokens *[is]*, *[of]*, *[from]*, respectively. At every iteration, we get an extraction by labeling the words using a fully-connected layer. Embeddings of the generated labels are added to the iterative layer embeddings. 45
4.4 P-R curve of IMoJIE, Gen2OIE, CIGL and CIGL with generation rescoring. . . 50
4.5 P-R curve of IMoJIE with no rescoring, label rescoring and generation rescoring. 50
4.6 P-R curve of CIGL with no rescoring, label rescoring and generation rescoring. 50
4.7 P-R curve of Gen2OIE with no rescoring, label rescoring and generation rescoring. 51

5.1 IGL-CA identifies conjunct boundaries by labeling a 2-D grid. This generates simple sentences, and CIGL-OIE emits the final extractions. 55
5.2 Process for manual comparison. Each extraction from both systems is presented to the annotator in a randomized order. The annotator checks if the extraction can be inferred from the original sentence and marks it accordingly. 59
5.3 MTGEN (multi-task Seq2Seq model) classifies the example into (non) compositional classes and generates the interpretation where valid, while UNIGEN (unified generation model), uses a Seq2Seq model to generate interpretations or identify non-compositional examples using a specific string "is not compositional". 64
5.4 The plot of relation distribution in the PRONCI dataset. It shows the number of relations that have a frequency of 1 to 9 and >=10. 66
5.5 Open IE Pipeline. Postprocessing of the extraction integrated with noun compound interpretation generates the new extraction. 74
5.6 Flowchart of the OpenIE-6.2 system. It allows flexibility of choosing from three Open IE systems (IMoJIE, Gen2OIE, CIGL), adding two linguistic features (Coordination Structures, Noun Compounds) and rescoring using two models (Labeling, Generative) 75

6.1 Crosslingual Data Transfer pipeline from English to Spanish. Firstly, The sentence and ext-sentences in English are aligned with a translation of the sentence (**Source Sentence** + **Translated Sentence** → **Aligned Sentence** and **Source Ext-sentence** + **Translated Sentence** → **Aligned Ext-sentence**). Secondly, the AACTRANS model uses the aligned text to generate the final consistent translations (**Aligned Sentence** → **Target Sentence** and **Aligned Ext-Sentence** → **Target Ext-Sentence**). Finally, Cross Lingual Projection (CLP) introduces **S**, **R**, **O** tags in the extraction (**Target Ext-Sentence** + **Input Extraction** → **Target Extraction**). 78

7.1 Distribution of languages of fact surface forms (in millions) on a subset of Wikidata. Compared to English and a few other languages, fact surface forms in Indian languages (the last five: HI, TE, TA, UR, GU) are extremely sparsely represented. 87
7.2 REFCOG architecture for linking Hindi sentences with KG facts (using their English surface forms). Fact-Text Dual Encoder scores the text, *T*, with all the KG facts, F_i, and outputs the top-k facts. A generative Seq2Seq model encodes the text *T* concatenated with top-k retrieved facts. A constrained decoder is then used to generate the correct fact. 91

7.3 Independent and Joint Classification for re-ranking the facts output by the retrieval model. 92
7.4 The two stage architecture. Stage 1 model outputs top-k entities that the Stage 2 model uses to generate contextual entity embeddings. The embeddings are passed through an MLP to get the final score for each entity. 98

List of Tables

2.1 Mapped continuous phrases between English (E) and Spanish (S) language sentences from the phrase extract algorithm 23

3.1 IMoJIE vs. CopyAttention. CopyAttention suffers from stuttering, which IMoJIE does not. 31
3.2 IMoJIE vs. OpenIE-4. Pipeline nature of OpenIE-4 can get confused by long convoluted sentences, but IMoJIE responds gracefully. 31
3.3 Comparison of various Open IE systems: non-neural, neural and our proposed models. Gen2OIE outperforms all other systems. (*) Cannot compute AUC because Sense-OIE and MinIE do not emit confidence values for extractions, and released code for Span-OIE does not include calculation of confidence values. 36
3.4 Performance of models that attempt to address the redundancy issue prevalent in generative neural Open IE systems. All systems are bootstrapped on OpenIE-4. 39
3.5 Measuring redundancy of extractions. MNO stands for Mean Number of Occurrences. IOU stands for Intersection over Union. 39
3.6 Performance of IMoJIE and GenOIE architectures with BERT/LSTM and T5 base architectures. IMoJIE achieves similar performance with either of the architectures, but GenOIE achieves a significant increase. However, at the higher performance levels of IMoJIE, LSTM seems to better at confidence scoring compared to the transformer-based T5, resulting in a 1.5% drop in AUC from 33.1 to 31.6. 40
3.7 Performance and Speed of labeling Open IE systems (RnnOIE, Multi^{2}OIE) and generative Open IE systems (IMoJIE, GenOIE, Gen2OIE) evaluated on the CaRB benchmark. Generative systems lead to better performance at the cost of slower inference speeds. 42

4.1 For the given sentence, IGL based Open IE extractor produces an incomplete extraction. Constraints improve the recall by covering the remaining words. . . 43
4.2 Evaluation of Open IE. Using constrained learning, CIGL-OIE gives better F1 than IMoJIE and reaches close to Gen2OIE. MinIE, SenseOIE, SpanOIE do not output confidence. The code of SenseOIE is not available to compute speed. *For RnnOIE, the reported speed is 149.2 sentences/sec, however, we have only been able to reproduce 64 sentences/sec with their latest implementation. . . . 48
4.3 The F1 and AUC scores of the three models – IMoJIE, CIGL and Gen2OIE using the original model confidence, generation rescoring and label rescoring. . 51
4.4 Performance and the number of constraint violations for training with different sets of constraints. CIGL-OIE represents training IGL architecture-based Open IE extractor with all the constraints: POSC, HVC, HVE and EC. 51

4.5 Evaluation of IMoJIE, Gen2OIE, IGL-OIE and CIGL-OIE using different metrics proposed for Open IE. 52

5.1 For the given sentence, IGL based Open IE extractor produces an incomplete extraction. Constraints improve recall by covering the remaining words. Coordination Analyzer handles hierarchical conjunctions. 54
5.2 P, R, F1 of the system evaluated on Penn Tree Bank for different systems. We use both BERT-Base and BERT-Large as the encoder 56
5.3 Evaluation of CaRB and CaRB(1-1) on two sentences. CaRB under-penalizes Open IE systems for incorrect coordination split by giving a recall of 100% for the second example of System 2. On the other hand, CaRB(1-1) reports the recall as 50% in the second example for System 2. 56
5.4 Wire57 F1 scores of IMoJIE and CIGL-OIE with addition of different coordination analyzers. IGL-CA improves both of the Open IE extractors. 58
5.5 Manual comparison of Precision and Yield on 100 random conjunctive sentences from CaRB Gold. 58
5.6 Adding a coordination analyzer, IGL-CA, to IMoJIE, Gen2OIE and CIGL, improves the score consistently in the CaRB(1-1) metric that is suitable for evaluating conjunctive sentences. Label rescoring is consistently used in all the experiments. 59
5.7 Examples of common and proper noun compounds along with their semantic interpretations ("*;*" separates multiple interpretations). [NON-CMP] indicates the absence of implicit relation between the constituent nouns. 60
5.8 Instructions for the task along with examples and common pitfalls that are provided to the human workers from AMT for constructing PRONCI dataset. 63
5.9 Examples demonstrating the addition of different sources of knowledge for the compound, "Buddhist monks", in form of prompts that are concatenated with [*SEP*] token. NNP and NN correspond for information about proper and common nouns respectively, which can be from WordNet, Named Entity tags or Wikipedia. 64
5.10 The number of training, validation and testing examples in the PRONCI dataset. CMP indicates the subset that contains only compositional examples and constitutes 63.9% of the examples. Non-CMP indicates the complementary subset that contains only non-compositional examples and constitutes the remaining 36.1% of the examples. 66
5.11 Performance of MTGEN and UNIGEN on the PRONCI dataset trained under five different knowledge settings. All the models are evaluated using the three types of matching. 'None' corresponds to using no external knowledge. Adding external knowledge improves the performance of the models in three out of four cases. 68
5.12 Performance of T5 model without any finetuning. Ponkiya et al. (2020) corresponds to the zero-shot setting adapted from the corresponding paper. Few-shot techniques use either five or ten example demonstrations. In 'Rand' the few-shot examples are chosen randomly while in 'KNN' the nearest neighbours of the query are chosen as the few-shot examples. Availability of annotated examples from PRONCI helps to substantially improve the performance of the model. Overall performance remains inferior to the finetuned models. 69

5.13 UniGen evaluated after random shuffling of characters in the proper (NNP) or common (NN) noun. 70
5.14 Quality of metrics evaluated using Pearson and Kendall rank correlation. (tuned) indicates models that are fine-tuned on 500 manually evaluated comparisons. . 71
5.15 Performance of the two models, MtGen and UniGen on the randomly split ProNCI dataset trained under five different knowledge settings. 71
5.16 Performance of T5 model without any finetuning on the random split of ProNCI dataset. 71
5.17 Performance of the UniGen model on the ProNCI dataset trained using different initializations of the Seq2Seq model. Random initialization leads to a huge drop in performance. 72
5.18 Performance of the UniGen model on ProNCI dataset trained with additional sources of knowledge added over Sentence knowledge. The additional sources do not provide further benefits. 72

6.1 Open IE examples transferred from English to Spanish, using both Independent (Indp) and Consistent (Const) translations. Independent translation results in inconsistencies which may have the same meaning (by using synonyms, fallecido vs. caído) or may change the meaning (changing gender from male to female, moderno to moderna). Consistent translation avoids these issues, resulting in better-quality training data. 77
6.2 Data statistics for Open IE examples and (English, language *F*) parallel sentences. 81
6.3 F1 and AUC performance of Open IE systems in Spanish (ES), Portuguese (PT), Chinese (ZH), Hindi (HI) and Telugu (TE). Training with AACTrans+CLP data shows strong performance with both GenOIE and Gen2OIE models. We also report the results of training Gen2OIE model with mT5 on all languages. . 82
6.4 Ablations of Gen2OIE model trained with AACTrans+CLP data on ES, ZH and HI. We analyze the effect of removing three components and re-training the model: 1. Sentence Consistency used in AACTrans data generation, and 2. Relation Ordering is used, and 3. Relation Coverage used in Stage-1 model training. 83
6.5 Evaluating inconsistency between translated extractions and corresponding sentences. 83
6.6 Evaluating CaRB F1 and AG of Gen2OIE predictions trained on SentExtTrans+CLP and AACTrans+CLP data. We find a decreasing trend of AG with increasing F1. 84
6.7 BLEU scores of translation and AAC-translation are similar showing that the performance improvement is because of the added consistency. 84
6.8 Unsupervised alignment perplexity for mBERT (MA) and Trained (TA) aligners 85
6.9 F1 and AUC of Gen2OIE trained with examples generated using TA and MA alignment strategies. (1, 2) corresponds to aligner 1 being used in AACTrans and aligner 2 being used in CLP. 85

7.1 KB linking task examples. Multilingual fact linking involves discovering the subset of KB facts expressed in a sentence, even when fact labels are available in a different language, requiring cross-lingual inference (Hindi-English in the above example). Fact-linking systems only output facts already present in the KB. Canonical fact extraction aims to discover new canonical facts not present in the KB while using the entities and relations defined in the existing KB schema. In contrast, Open IE extracts open-ended facts that may or may not correspond to entities, relations, or facts defined in the KB. *Q* and *P* represent the entity and property identifiers in Wikidata. The fact identifiers (e.g., F_{23}) are assigned and are not part of Wikidata. 88
7.2 The new INDICLINK dataset (Section 7.1.3) contains examples in English and corresponding manually translated test examples in six Indian languages. KG fact surface forms are always available in English but are only sparsely available in other languages. 90
7.3 Comparison of different models on the INDICLINK dataset. REFCOG with ALL-Sum dual encoder and EL cross encoder, outperforms independent (INDCLS) and joint (JNTCLS) classification based re-ranking on top of $DE_{ALL\text{-}Sum}$. Ablations indicate the importance of DE and joint prediction of S, R and O for the REFCOG model. Constraints reduce the P@1, R@5 metrics but ensure production of only valid facts. Please see Section 7.1.6.1 and Section 7.1.6.2 for further details. . . 94
7.4 Multilingual fact surface forms in Retrieval and Generation models (Section 7.1.6.3). EL, TL, ETL and ALL correspond to descriptions in English, language of input text *T*, EL+TL and all languages, respectively. Concat, Max and Sum refer to concatenation, max and sum scoring operations. For REFCOG, we use ALL-Sum facts for retrieval and experiment with different fact surface forms for cross-encoder. 96
7.5 P@1, macroP@1 of REFCOG with fact surface forms in various languages at cross encoder stage. The macroP@1 is evaluated for the Complete test set as well as the Subset where descriptions are available in all languages. Improvement in macroP@1, indicates stronger performance on facts with less-frequently occurring relations. 96
7.6 Evaluation of KG facts linked to Open IE extractions. 97
7.7 Statistics of the dataset used. 100
7.8 Link Prediction performance on OLPBENCH. 101
7.9 H@1 with increasing top-*k* Stage-1 samples. 101
7.10 Ablation of the best CEAR model, which shows the importance of BERT pre-trained knowledge, Cross-Entity Attention and Stage-1 Entity Ranks. 101