It’s My Job:
Identifying and Improving Content Quality for Online Recruitment Activities

Nidhi Goyal
Ph. D Scholar
IIIT-Delhi

INDRAPRASTHA INSTITUTE *of* INFORMATION TECHNOLOGY **DELHI**

Evaluation Committee

Thanks to the Committee Members

Dr. Charu Sharma
IIIT-Hyderabad

Ph.D. Advisor
Prof. Ponnurangam Kumaraguru ("PK")
IIIT-Hyderabad

Dr. Rajiv Ratn Shah
IIIT-Delhi

Ph.D. Co- advisor
Dr. Niharika Sachdeva
InfoEdge India Limited

Dr. Arun Balaji Buduru
IIIT-Delhi

Ph.D. Co- advisor
Dr. V. Raghava Mutharaju
IIIT-Delhi

Outline

- Online Recruitment Ecosystem
- Motivation
- Problem Statement
- Research Mission
- Literature review
- Contributions
- Summary
- Timeline
- Publications
- References

Online Recruitment Ecosystem

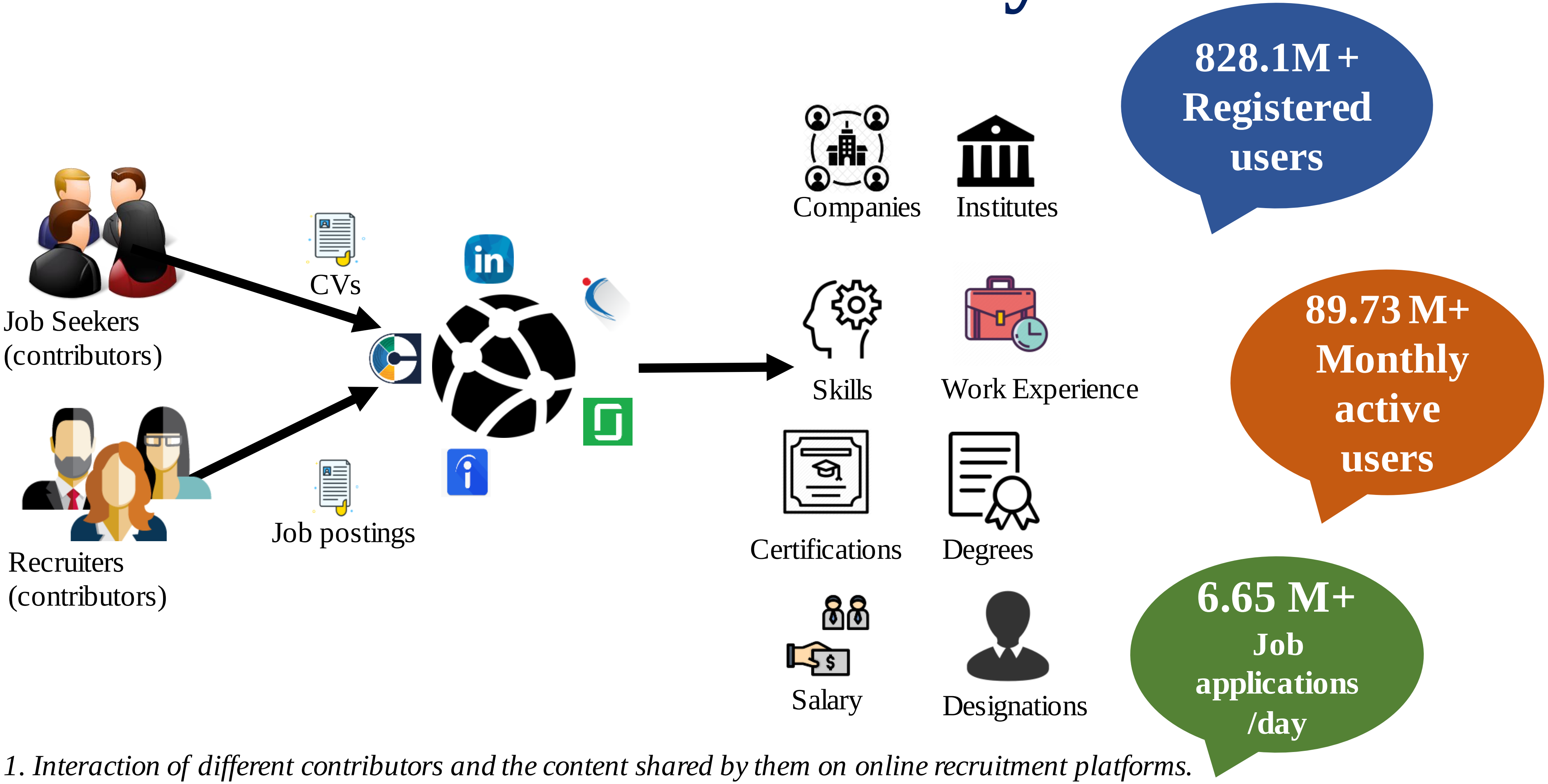


Fig. 1. Interaction of different contributors and the content shared by them on online recruitment platforms.

Statistics Source: https://datareportal.com/reports/digital-2022-april-global-statshot

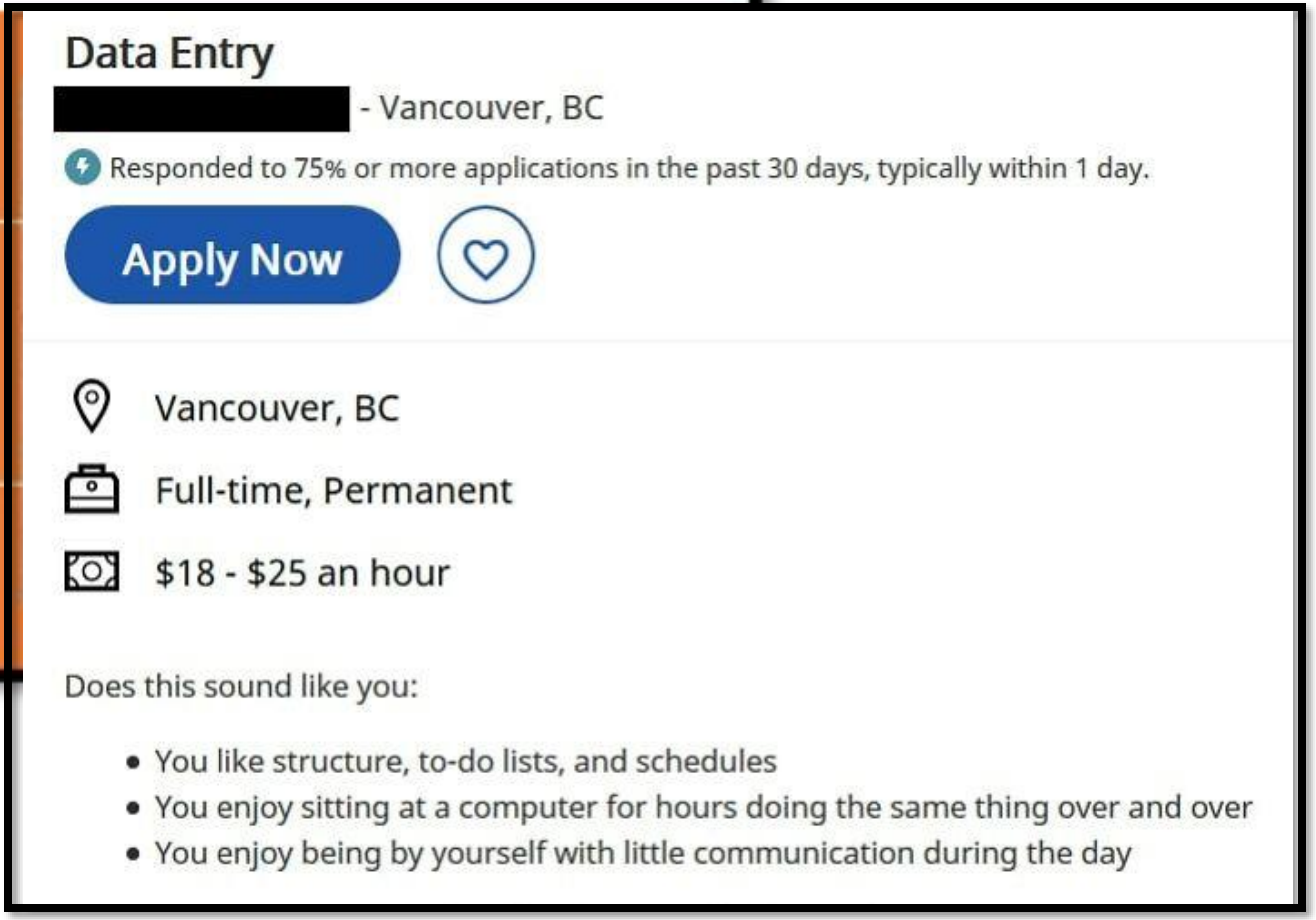
Motivation


LINKEDIN PHISHING SCAM: HACKERS TARGET USERS WITH FAKE JOB OFFERS
By Prashant Tilekar | 28 May 2021 | 4 min read
0 Shares
Quick Heal
Security Simplified
Scam Alert:
Beware of Fake LinkedIn Job Offers


Watch Out for Scammers When Job Hunting
FTC cracking down on companies suspected of employment fraud
by Kenneth Terrell, AARP, February 20, 2020
JOB SEARCH
FIND JOB
Man searching for a hob
ALEKSANDR DAVYDOV / ALAMY STOCK PHOTO

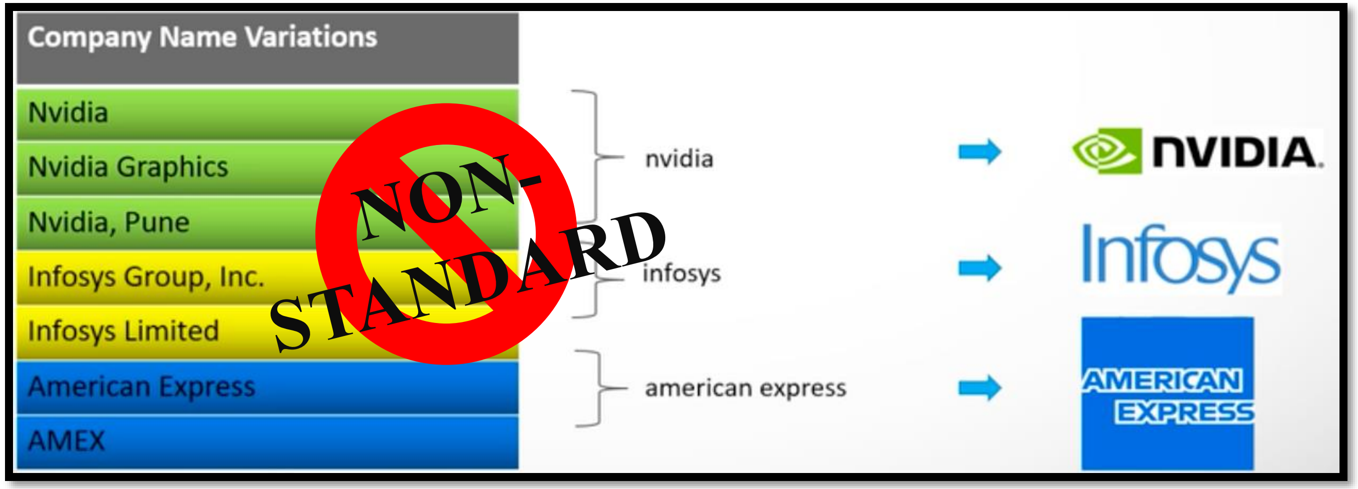
Motivation

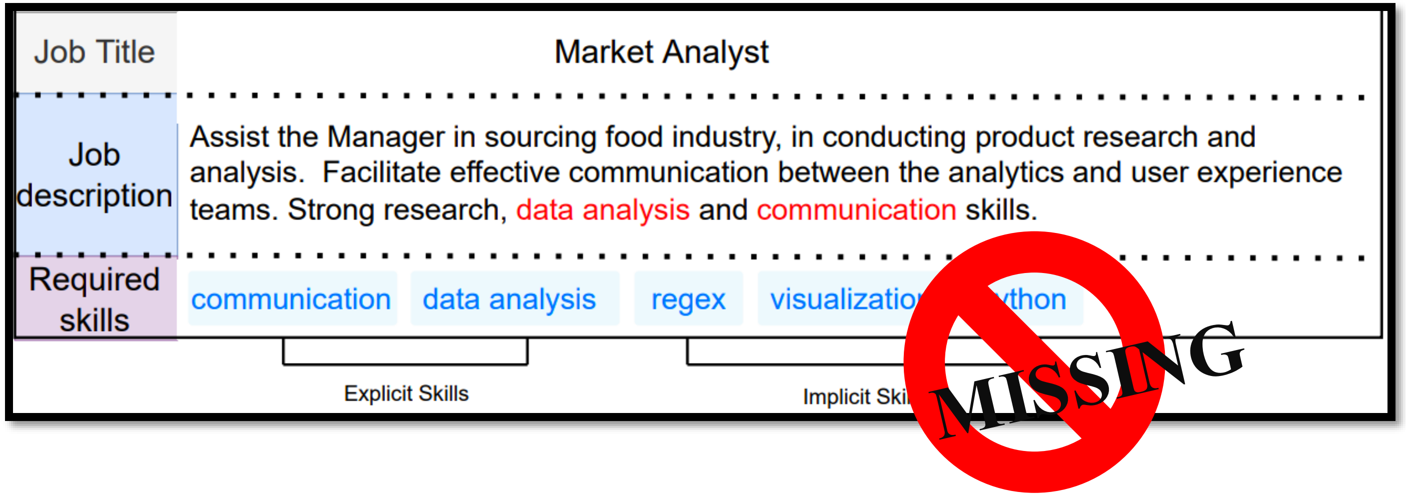

How to write a job posting that stands out?
cloudely
HOW TO WRITE
A JOB POSTING
THAT STANDS OUT?


Data Entry
[redacted] - Vancouver, BC
Responded to 75% or more applications in the past 30 days, typically within 1 day.
Apply Now
Vancouver, BC
Full-time, Permanent
$18 - $25 an hour
Does this sound like you:
• You like structure, to-do lists, and schedules
• You enjoy sitting at a computer for hours doing the same thing over and over
• You enjoy being by yourself with little communication during the day

Problem Statement

Data Entry Clerks Position
We have several openings available in this area earning $1000.00-$2500.00 per week. We are seeking only honest, self-motivated people with a desire to work in the home typing and data entry field, from the comfort of their own homes.The preferred applicants should be at least 18 years old with Internet access. No experience is needed. However the following skills are desirable: Basic computer and typing skills, ability to spell and print neatly, ability to follow directions.
Earn as much as you can from the comfort of your home typing and doing data entry.
You do NOT need any special skills to get started.





Research Mission

Identifying misleading, non-standard, missing content and **improving** content quality on online recruitment platforms by leveraging domain-specific knowledge and deep learning-based approaches.

What to Identify?

What to Improve?

How to Identify and Improve?

What to Identify?

Role : Java Developer/Senior Developer/Architect - Spring Boot/Microservices Architecture
Job Requirements : Java Microservices, Application Deployment, Application High-Level Design
UG : B.Tech/B.E. in Any Specialization, B.Sc in Any Specialization, BCA in Any Specialization

Facts →

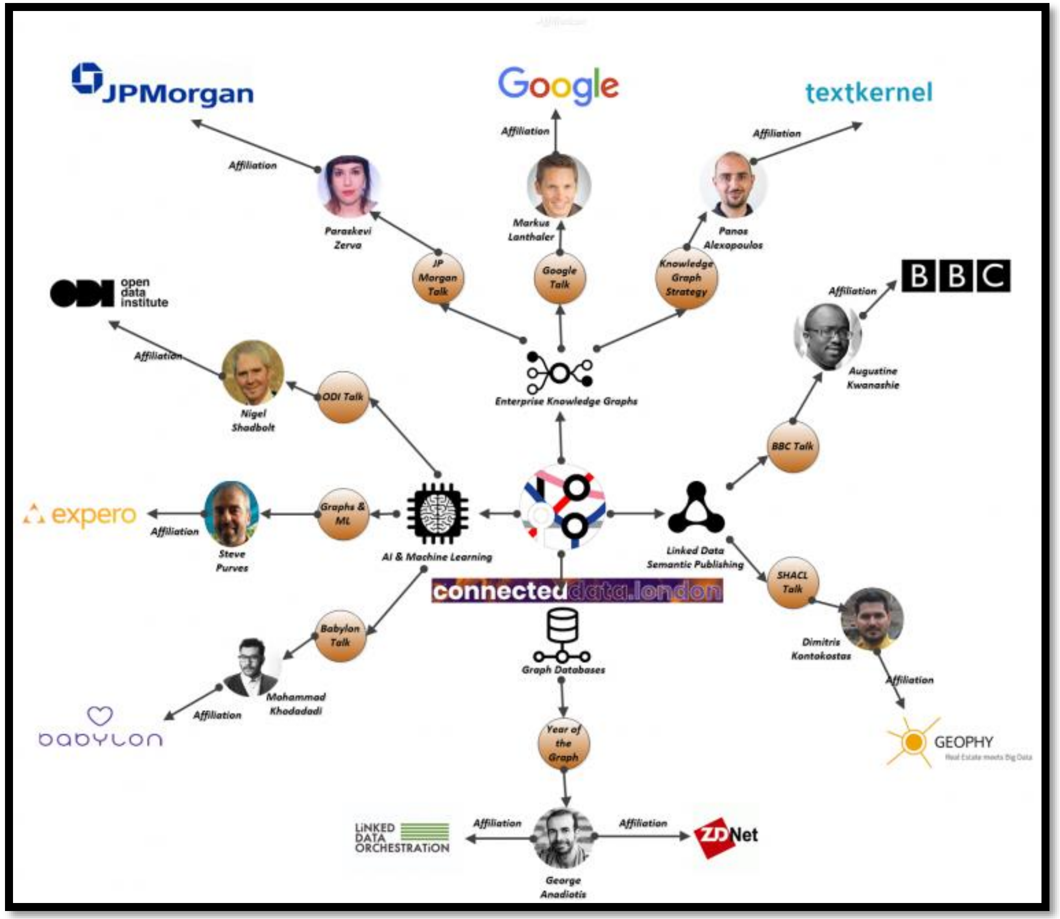
Entities (Java Microservices, Architect, etc.)

(Java Microservices, is a, skill)
(B. Tech, is a, Degree)
(B. Sc, is a, Degree)
(BCA, is a , Degree)
(Java Developer, is a , designation)
(Senior Developer, is a , designation)
(Architect, is a , designation)
(Spring Boot, is a , skill)

Entity types
(Skill, Designations, Degree)

How to **Identify** facts ?

Extract this information from unstructured text
Convert to structured format (what)?



Literature

Existing Knowledge Graphs

Facebook's Entity Graph

yago select knowledge

Freebase

DBpedia

Identification

Information Extraction

- Generic frameworks
- Domain-specific methods

Kertkeidkachorn et al. (AAAI 2017)

Wang et al. (CIKM 2018)

- Content-based approaches
- Context-based approaches
- Knowledge-based approaches

Vidros et al. (Future Internet 2017)

Mahbub et al. (ISD 2018)

Pan et al. (ISWC 2018)

Research Gap

Existing Knowledge Graphs

Facebook's Entity Graph

yago select knowledge

Freebase

DBpedia

Identification

Information Extraction

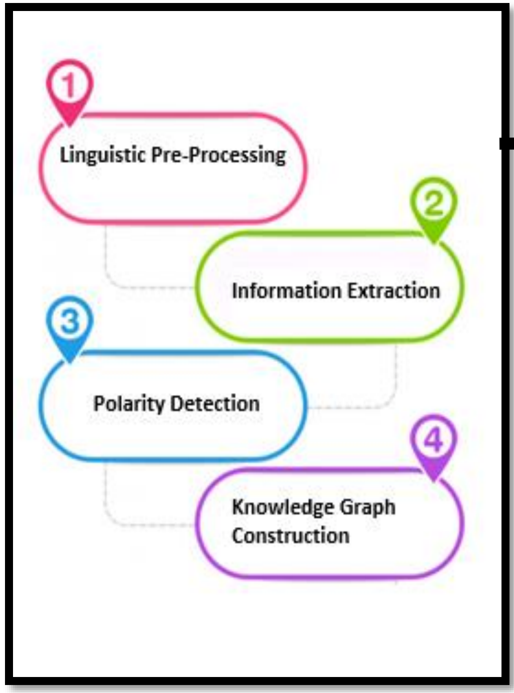
- Generic frameworks
- Domain-specific methods

Kertkeidkachorn et al. (AAAI 2017)

Wang et al. (CIKM 2018)

These methods/ KGs are specific to general concepts and lack domain-specific facts, important entities such as evolving skills, designations, and hidden properties of job such as type of recruiter, shift timings, etc.

Contribution 1: Building the Domain-Specific Knowledge Graphs



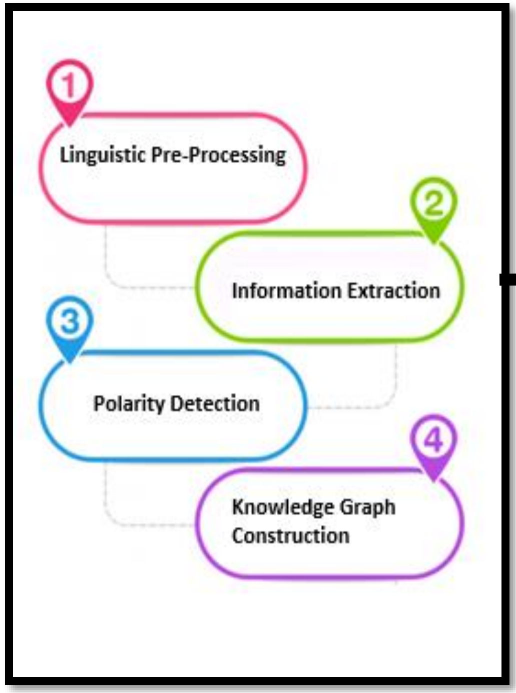
- Preprocess the noisy, unstructured and semi-structured data from job postings using NLP techniques

- To accomplish this task, we
  - Employed **sentence detection module**
  - Revived **missing phrases** using POS Tagging
  - Removed **HTML Non-ASCII** characters.

- Exploit rule -based heuristics and vocabulary list to deal with **Abbreviations**

Con2KG-A Large-scale Domain-Specific Knowledge Graph published in Proceedings of the 30th ACM Conference on Hypertext and Social Media, pp. 287-288. 2019.

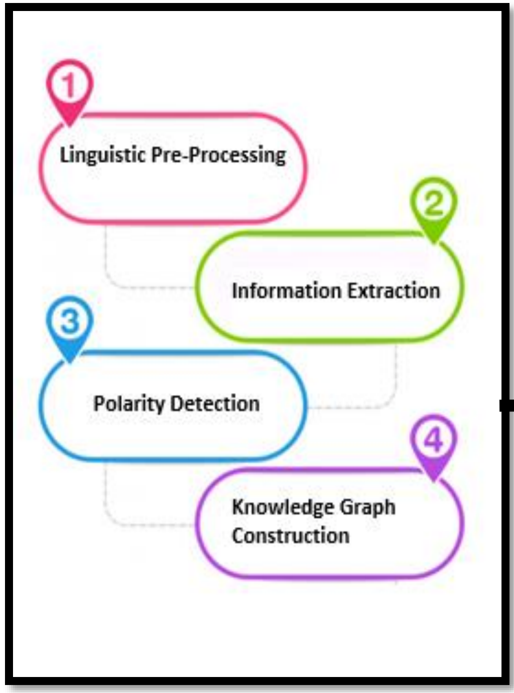
Contribution 1: Building the Domain-Specific Knowledge Graphs

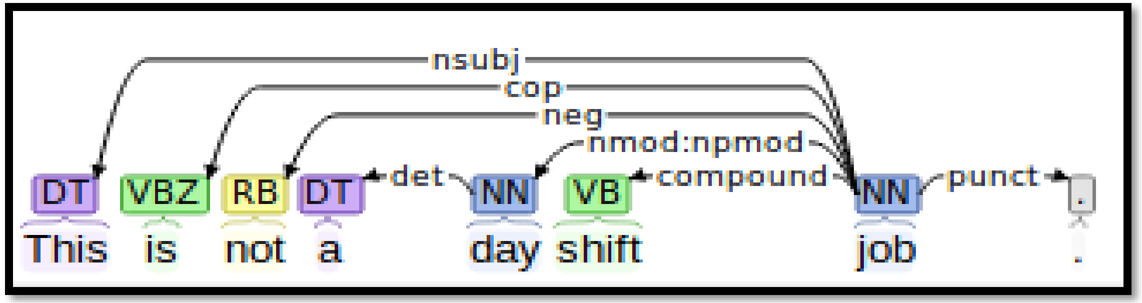


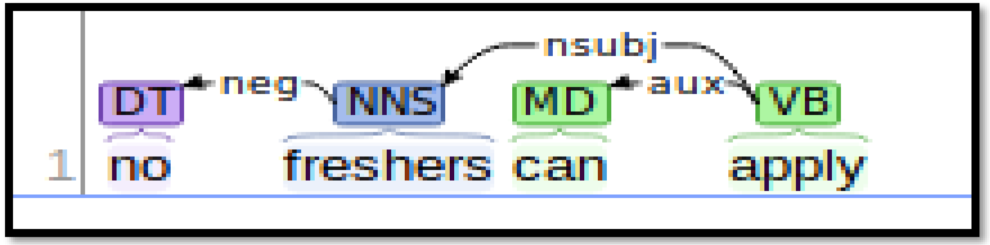
- POS tagging and NER using **Stanford NLP, spacy, FlashText, and customized libraries**.
- Dependency Parsing to find the context.
- **Relation and triple extraction** using OpenIE Systems.

Con2KG-A Large-scale Domain-Specific Knowledge Graph published in Proceedings of the 30th ACM Conference on Hypertext and Social Media, pp. 287-288. 2019.

Contribution 1: Building the Domain-Specific Knowledge Graphs





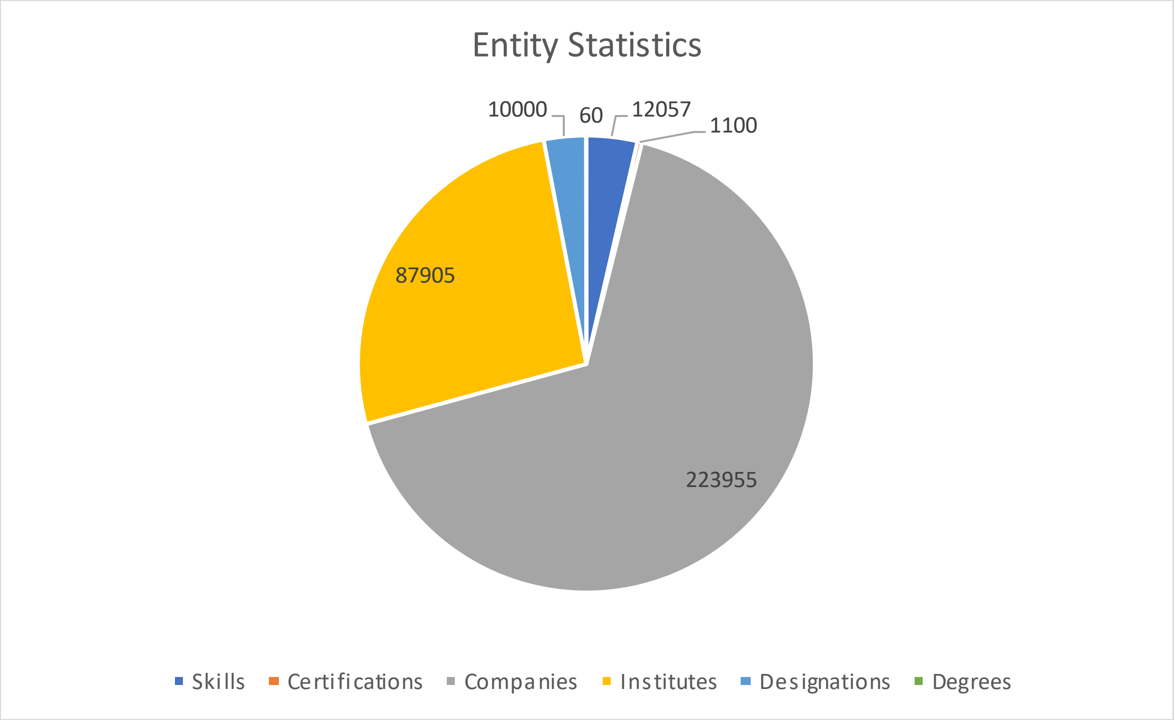


Dependency Parsing to tag entities with positive and negative polarities.

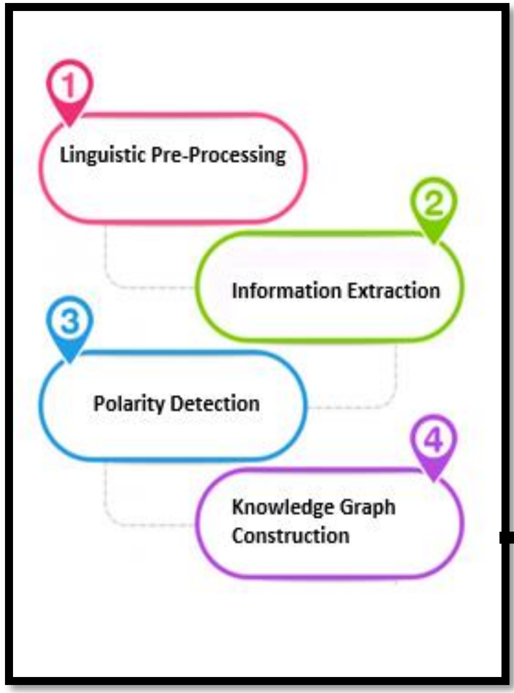
Con2KG-A Large-scale Domain-Specific Knowledge Graph published in Proceedings of the 30th ACM Conference on Hypertext and Social Media, pp. 287-288. 2019.

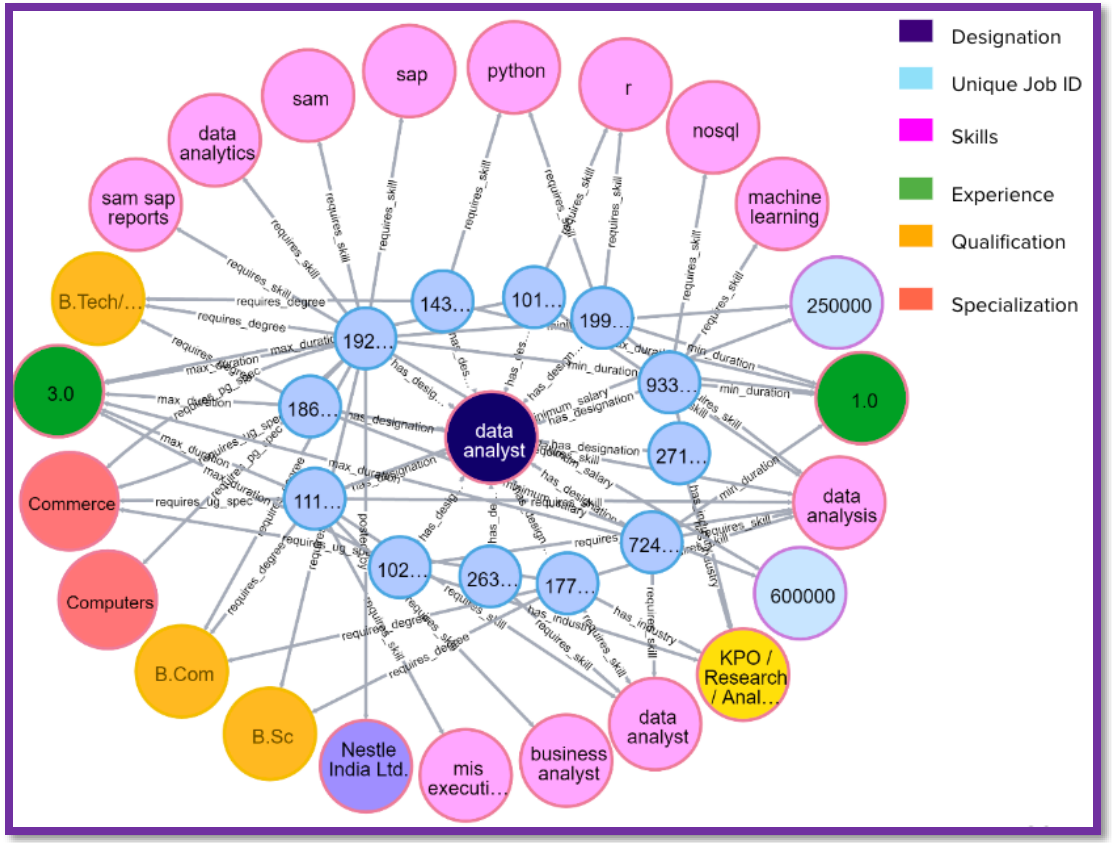
Contribution 1: Building the Domain-Specific Knowledge Graphs

- 250,000 Job postings
- 5,220 unique relations linking 3,65,0,61 entities
- 40,11,030 relationships

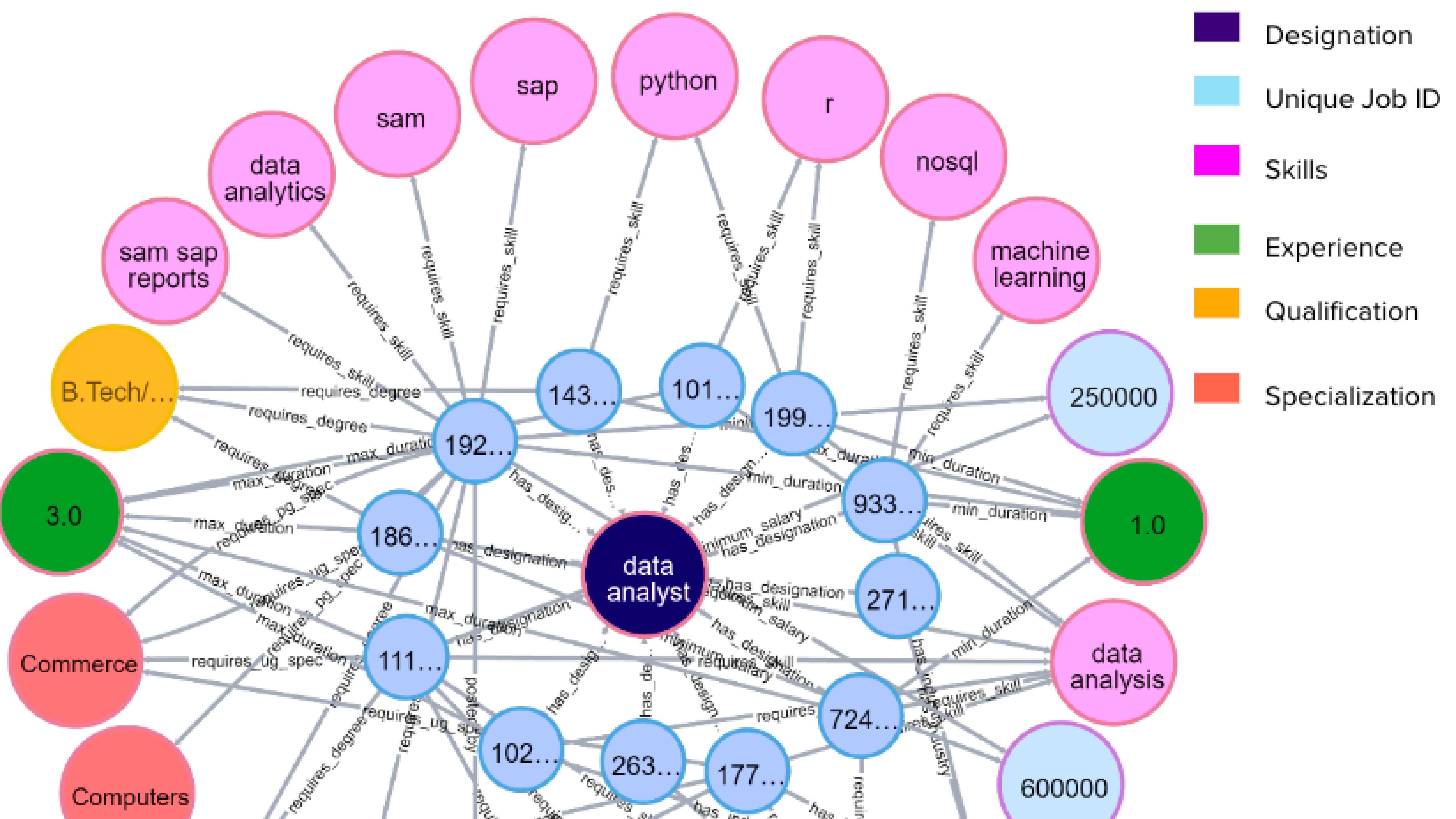


Contribution 1: Building the Domain-Specific Knowledge Graphs





Con2KG-A Large-scale Domain-Specific Knowledge Graph published in Proceedings of the 30th ACM Conference on Hypertext and Social Media, pp. 287-288. 2019.



Designation
Unique Job ID
Skills
Experience
Qualification
Specialization
sam sap reports
data analytics
sam
sap
python
r
nosql
machine learning
B.Tech/…
3.0
Commerce
Computers
250000
1.0
data analysis
600000
data analyst
143…
101…
199…
192…
186…
933…
271…
111…
102…
263…
177…
724…
requires_skill
requires_degree
max_duration
min_duration
has_designation
requires_ug_spec
requires_pg_spec
minimum_salary

Summary

- We randomly selected 310 jobs from our legacy dataset containing 4719 sentences to evaluate the quality and quantity of the triples.
- 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%.
- Triple extraction causes 0.05% errors due to incomplete triples.
- 0.20% due to no triple extraction for most of the sentences.

What to Identify?

- Fraudulent jobs contain untenable facts about domain-specific entities such as mismatch in skills, industries, offered compensation, etc.

Data Entry Clerks Position We have several openings available in this area earning $1000.00-$2500.00 per week. We are seeking only honest, self-motivated people with a desire to work in the home typing and data entry field, from the comfort of their own homes.The preferred applicants should be at least 18 years old with Internet access. No experience is needed. However the following skills are desirable: Basic computer and typing skills, ability to spell and print neatly, ability to follow directions. Earn as much as you can from the comfort of your home typing and doing data entry. You do NOT need any special skills to get started.	Data Entry Clerk Responsibilities include, but are not limited to: Review and process confidential and extremely time-sensitive applications. Identify objective data and enter (""key what you see"") at a high level of productivity and accuracy. Perform data entry task from a paper and/or document image. Utilize system functions to perform data look-up and validation. High volume sorting, analyzing, indexing, of insurance, legal and financial documents. Maintain high degree of quality control and validation of the completed work Identify, classify, and sort documents electronically.

Fig. 2. Examples of job postings a) fraudulent job on the left and b) legitimate at the right. These job postings are taken from publicly available dataset.

Literature

Identification

- Generic approaches
- Domain-specific methods

Kertkeidkachorn et al. (AAAI 2017)

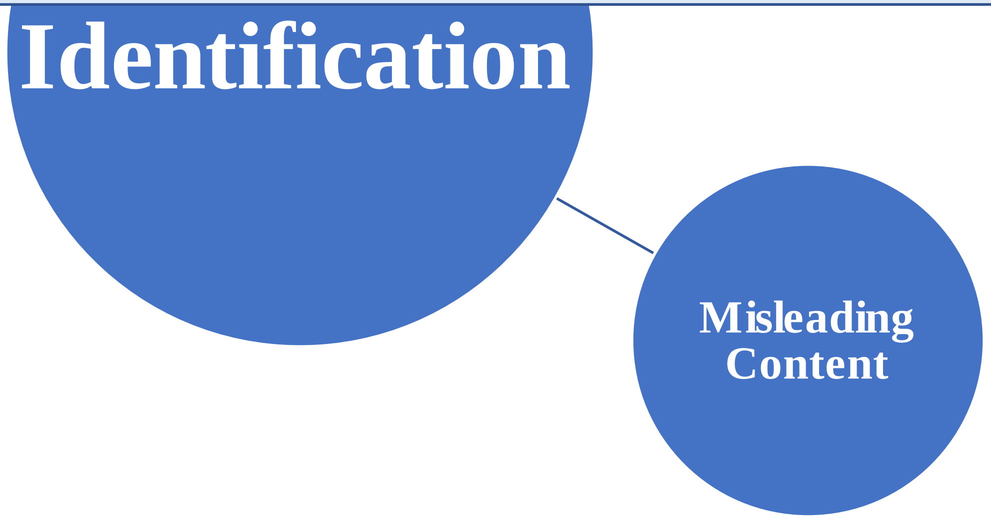
Wang et al. (CIKM 2018)

Misleading Content

- Content-based approaches → Vidros et al. (Future Internet 2017)
- Context-based approaches → Mahbub et al. (ISD 2018)
- Knowledge-based approaches → Pan et al. (ISWC 2018)

Research Gap

Handcrafted , linguistic, writing styles, string-based features.
Ignore the factual information among domain-specific entities present in job postings.



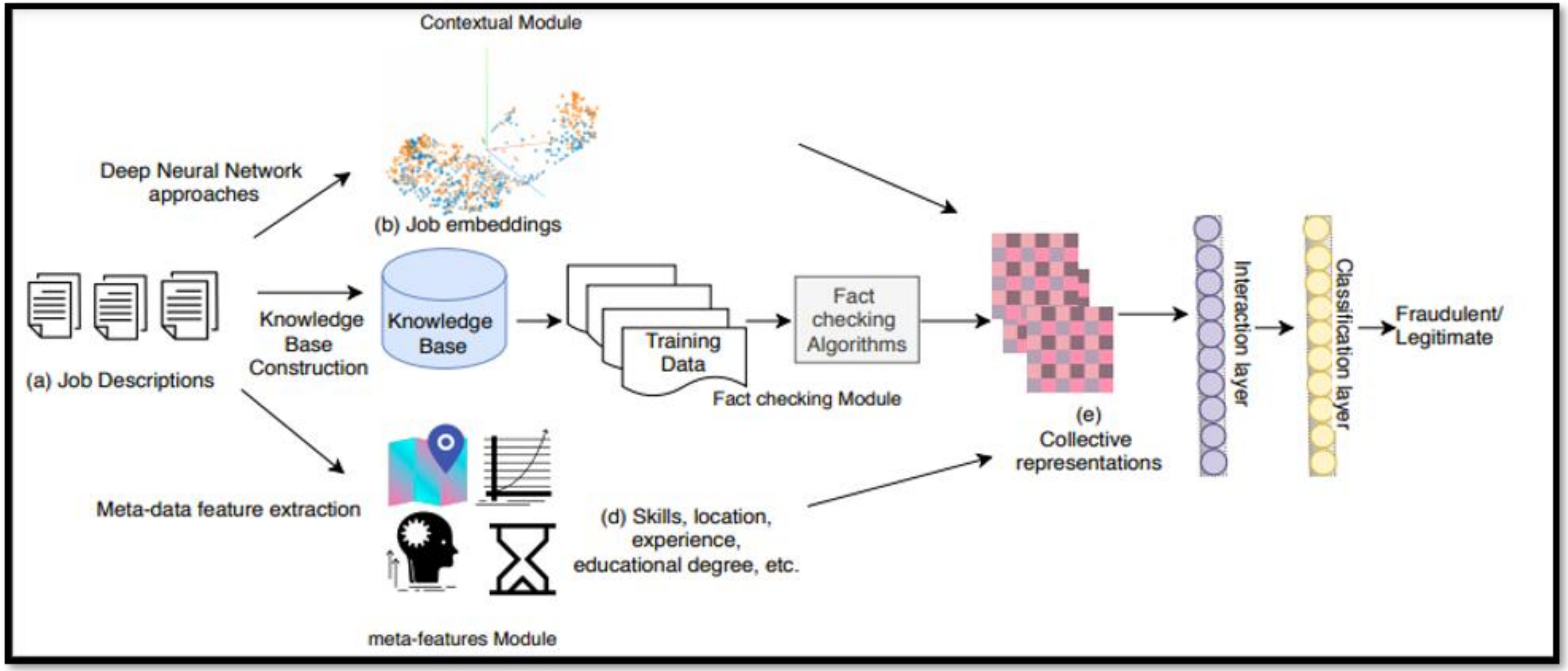
- Content-based approaches → Vidros et al. (Future Internet 2017)
- Context-based approaches → Mahbub et al. (ISD 2018)
- Knowledge-based approaches → Pan et al. (ISWC 2018)

Contribution 2: Identify misleading job postings using domain-knowledge

Our objective is to learn function ϕ where ϕ: F ($KG^{A}_{false}(T)^{i}$, $KG^{A}_{true}(T)^{i}$, c^{i}, m^{i}) where $KG^{A}_{true}(T)^{i}$ is the scoring function, we learn from triple $t^{i} \in T^{i} | y_i = 0$ of legitimate job postings and $KG^{A}_{false}(T)^{i}$ from triple $t^{i} \in T^{i}$ **$| y_i = 1$** of fraudulent job postings.

$KG^{A} \in$ {*TransE, TransR, TransH, TransD, DistMult, ComplEx, HolE, RotatE*}

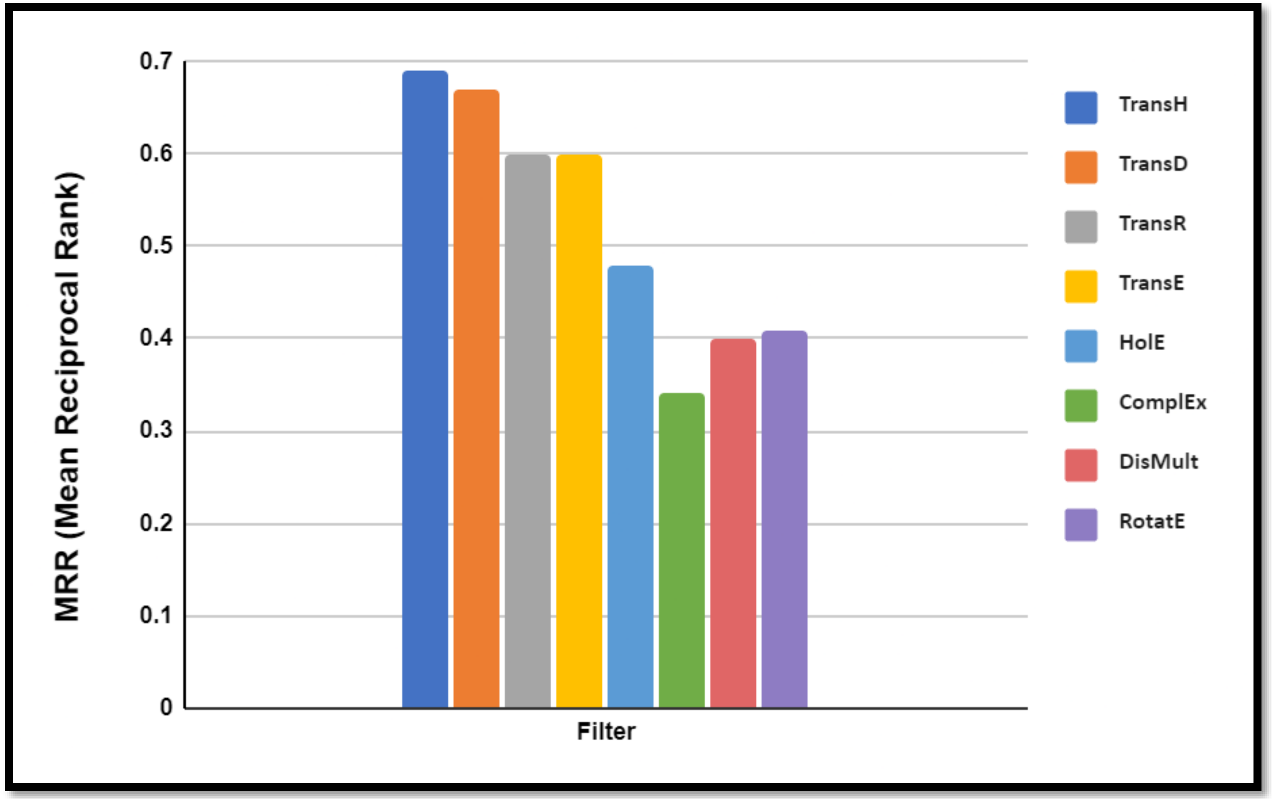
Contribution 2: Identify misleading job postings using domain-knowledge



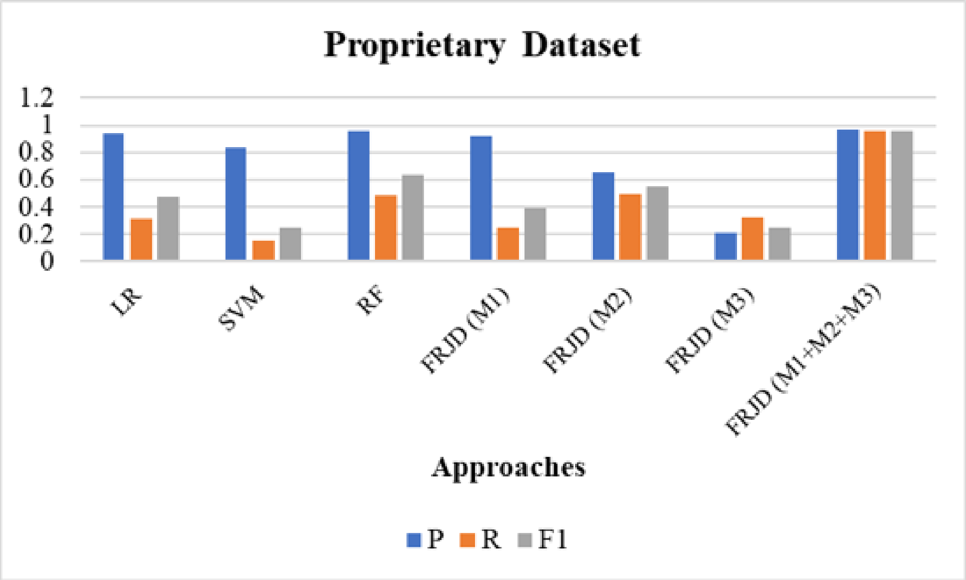
Spy The Lie: Fraudulent Jobs Detection in Recruitment Domain using Knowledge Graphs. Published in 14th International Conference on Knowledge Science, Engineering and Management (KSEM 2021).

Contribution 2: Identify misleading job postings using domain-knowledge

- MRR (Mean Reciprocal Rank) and Hits @n metrics for triple prediction where n={1,3,10}
- TransH outperforms the other fact-checking algorithms for our dataset.
- TransH is able to model many-to-many relationships well for our dataset.



Contribution 2: Identify misleading job postings using domain-knowledge



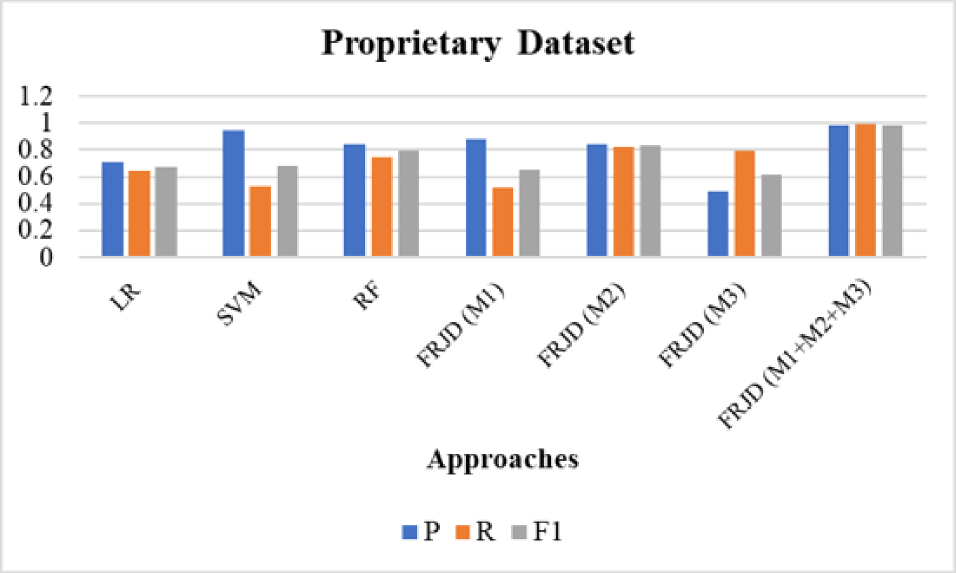
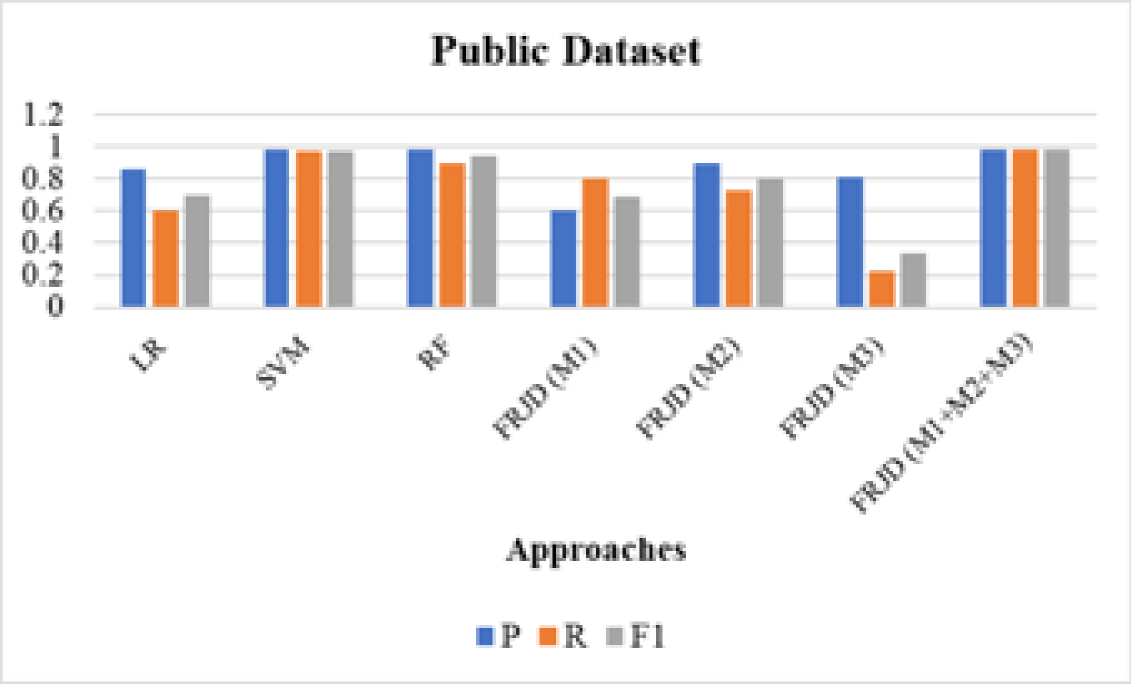


Fig. 3. Evaluation results on proprietary dataset for job postings a) fraudulent class and b) legitimate class at the right.

Contribution 2: Identify misleading job postings using domain-knowledge



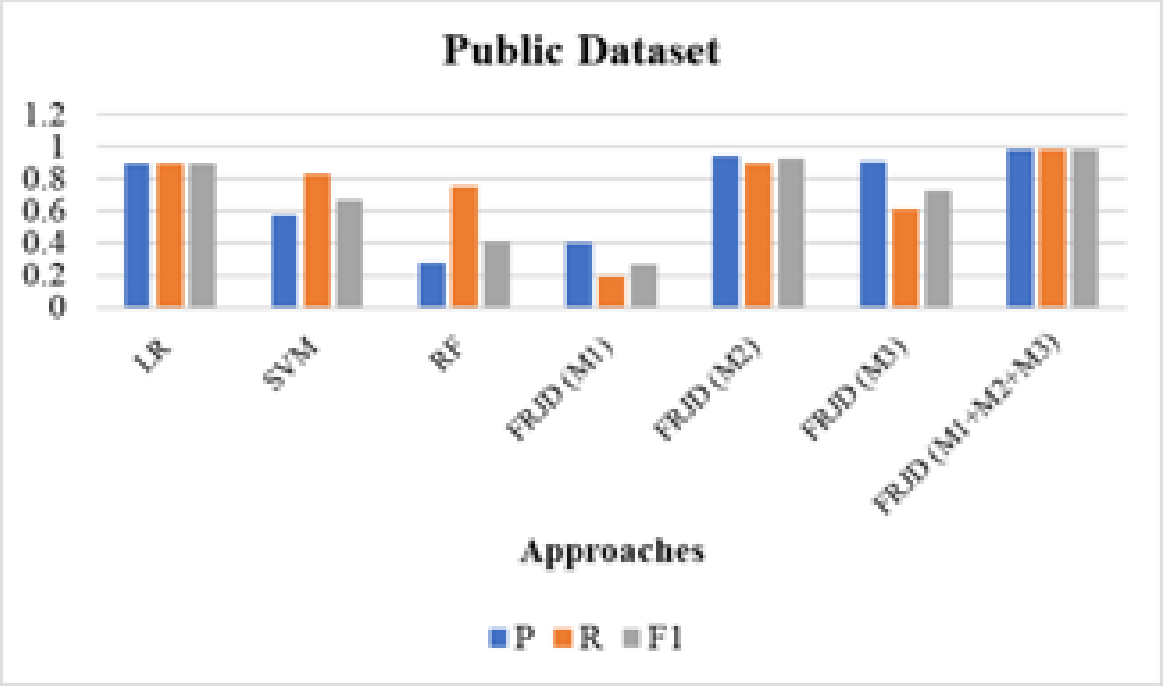


Fig. 3. Evaluation results on public dataset for job postings a) fraudulent class and b) legitimate class at the right.

Summary

- Study on a fact validation dataset containing 4 million facts extracted from job postings.
- Proposed a multi-tier novel end-to-end framework called **FR**audulent **J**obs **D**etection (FRJD), which jointly considers
- a) fact validation module using knowledge graphs,
- b) contextual module using deep neural networks
- c) meta-data inclusion

Spy The Lie: Fraudulent Jobs Detection in Recruitment Domain using Knowledge Graphs. Published in 14th International Conference on Knowledge Science, Engineering and Management (KSEM 2021).

What to Improve?

Recruitment Domain has non-standard user-generated entities everywhere !

ICICI Prudential Life Insurance has 497 variations

Dr. Babasaheb Ambedkar Marathwada University Aurangabad has 1145 variations

Senior Software Engineer has 123 variations

Microsoft Excel has 37 variations

What to Improve?

01	Spelling Variations	• Java Developer • Java Deveoper
02	Hierarchical variations	• Oracle Financial Services Software • Oracle Corporation
03	Overlapping but different entities	• Emerald Bikes pvt limited • Emerald Jewellery Retail Limited
04	Domain specific concepts	• SOAP • REST
05	Semantic variations	• Accel Frontline • Insiprisys
06	Short Forms	• umbc • University of Maryland, Baltimore

Research Gap

Canonicalization

- Generic approaches
- Domain-specific methods

Vashishtha et al. (WWW 2018)

Fatma et al. (PAKDD 2020)

Improving

Focus upon either statistical similarity measures or deep learning methods like word-embedding or siamese network-based representations for canonicalization.

Contribution 3: Improve quality of job postings

1

Unstructured text

User generated content is often noisy, ambiguous and contains duplicate information.

Company Profiles

Job postings

2

OpenKB

This leads to redundant information and increased KB size.

Knowledge Base

3

Performance

This affects performance in downstream tasks like question answering, search systems, recommendation etc.

Contribution 3: Improve quality of job postings

Consider E be the set of entities extracted from job postings, CVs, and company profiles. For each entity x_i, we consider its side information $s_i \in S \; \forall \; x_i \in E$ acquired from heterogeneous sources. Given two entities x_i and x_j and their corresponding side information s_i and s_j, we aim to find the mapping

$$F\,(x_i, s_i, x_j, s_j) \rightarrow \text{similarity}\,(x_i\,, x_j)$$

A pairwise similarity matrix (M_{sim}) is formed by applying F over the set of all entity pairs. A clustering algorithm is used to form unique canonical clusters of similar entities.

Contribution 3: Improve quality of job postings

Clustering / Canonicalizing

Hierarchical Agglomerative clustering.

KCNet

Kernel based Canonicalization Network to learn pairwise similarity between input pairs.

Dataset

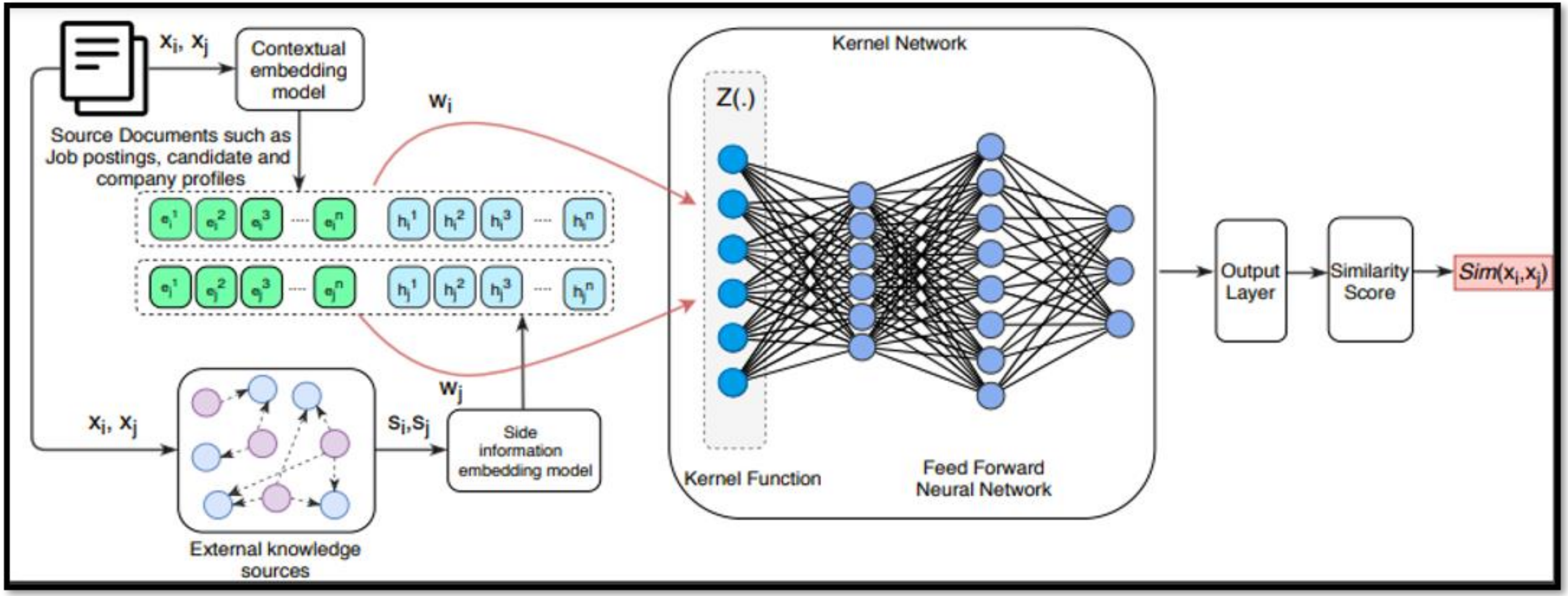
Entities extracted from documents.
Side Information Acquisition.
Positive and negative samples are created.
Random sampling is used for negative pairs.

KCNet: Kernel-based Canonicalization Network for entities in Recruitment Domain published in 30th International Conference on Artificial Neural Networks (ICANN). 2021.

Contribution 3: Improve quality of job postings

Source	Dataset	Entity Clusters
Proprietary	RDE(C)	25,602
	RDE(I)	23,690
	RDE(D)	3,894
	RDE(S)	607
Open	DBpedia(C)	2,944
	ESCO (S)	2,644
	ESCO (D)	2,903

Contribution 3: Improve quality of job postings



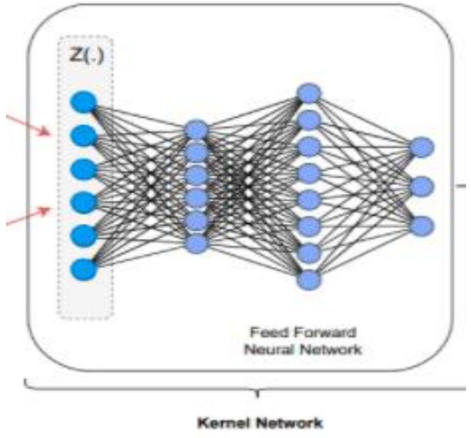
KCNet: Kernel-based Canonicalization Network for entities in Recruitment Domain published in 30th International Conference on Artificial Neural Networks (ICANN). 2021.

Contribution 3: Improve quality of job postings

- Z models element-wise relationships between input pairs.

$$Z = (w_i \circ w_j) \odot |w_i - w_j|$$

$$Z = \}\{w^1_i * w^1_j ,\ldots, w_i^{m+n} * w_j^{m+n}, |w^1_i - w^1_j| ,\ldots, | w_i^{m+n} - w_j^{m+n}|\}$$



Similarity(x_i, x_j)

where w^k_i represents the k^{th} dimension of w_i. The dimensionality of Z is $2*(m+n)$.

Contribution 3: Improve quality of job postings

Model	Performance							
	S		D		I		C	
	P	F		P	F	P	F	P
Galarraga-IDF†}	33.2	12.5	63.0	60.3	64.3	66.5	75.8	71.2
Distilled S-BERT(*)+cosine	47.8	47.5	49.7	48.8	49.7	49.1	49.2	49.1
Distilled S-BERT(**)+ cosine	47.5	48.8	49.8	49.9	34.6	41.5	56.2	48.4
CharBiLSTM+A†	81.8	86.9	72.6	77.2	84.5	84.8	99.3	98.9
WordBiLSTM+A†	80.1	86.5	90.5	94.8	80.6	83.3	95.3	95.6
CharBiLSTM+A+Word+A†	82.7	88.5	94.4	96.3	86.7	86.7	99.5	99.2
KCNet (without sideinfo)	96.7	90.6	99.6	90.9	92.4	89.3	99.4	98.8
KCNet (with sideinfo)	99.5	99.4	99.7	99.6	99.5	99.5	99.5	99.3

Table 1: Test Results of pairwise similarity using our proposed model in comparison with different baselines. Here S, D, I, C refers to Skills, Designations, Institutes, and Companies datasets (Proprietary) respectively. Results of † are taken from [1]. P and F refers to Precision and F1-scores. Distilled S-BERT (*, **) refers to (entity, entity side information) embedding using distilled S-BERT model.

Summary

- KCNet induces a non-linear mapping between the contextual vector representations while capturing fine-granular and high-dimensional relationships among vectors.
- KCNet efficiently models more prosperous semantic and meta side information from external knowledge towards exploring kernel features for canonicalizing entities in the recruitment domain.
- KCNet is able to model similar semantic variations *(mycology, fungi studies)* gives a pairwise similarity score of 0.98.
- Misclassified some skills such as *bees wax* and *natural wax* which signify same concept but occur in the different cluster.

Improving job postings quality by missing skills prediction (Work in progress)

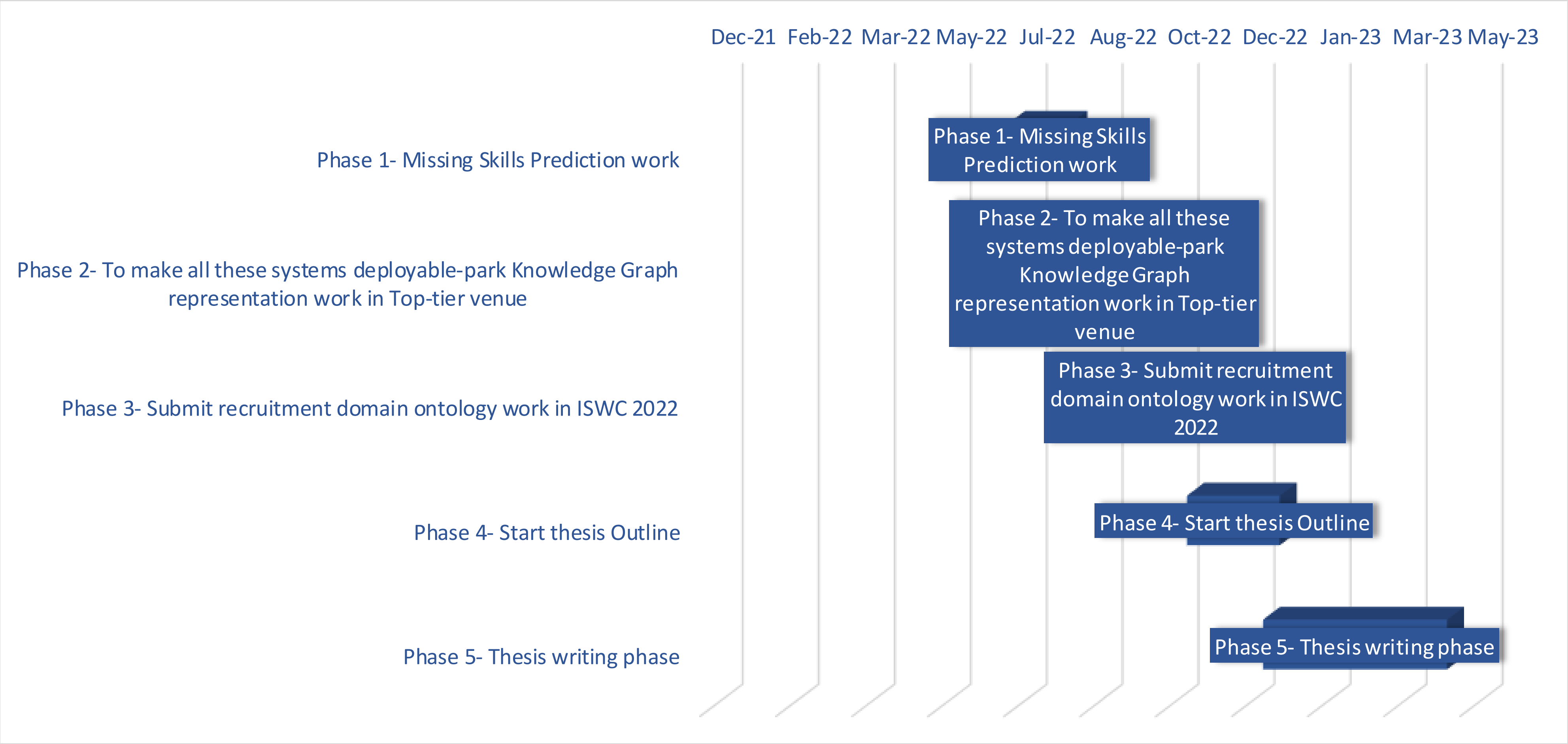
- Writing a good job posting is a crucial task
- poor quality jobs:
  - get less number of applies from job seekers
  - poor recommender systems performance
  - affect search systems

Skill is most important criteria

65% of Job postings miss relevant skills

40% of Job postings miss listing 20% or more explicitly-stated skills

Timeline



Dec-21
Feb-22
Mar-22
May-22
Jul-22
Aug-22
Oct-22
Dec-22
Jan-23
Mar-23
May-23
Phase 1- Missing Skills Prediction work
Phase 2- To make all these systems deployable-park Knowledge Graph representation work in Top-tier venue
Phase 3- Submit recruitment domain ontology work in ISWC 2022
Phase 4- Start thesis Outline
Phase 5- Thesis writing phase
Phase 1- Missing Skills Prediction work
Phase 2- To make all these systems deployable-park Knowledge Graph representation work in Top-tier venue
Phase 3- Submit recruitment domain ontology work in ISWC 2022
Phase 4- Start thesis Outline
Phase 5- Thesis writing phase

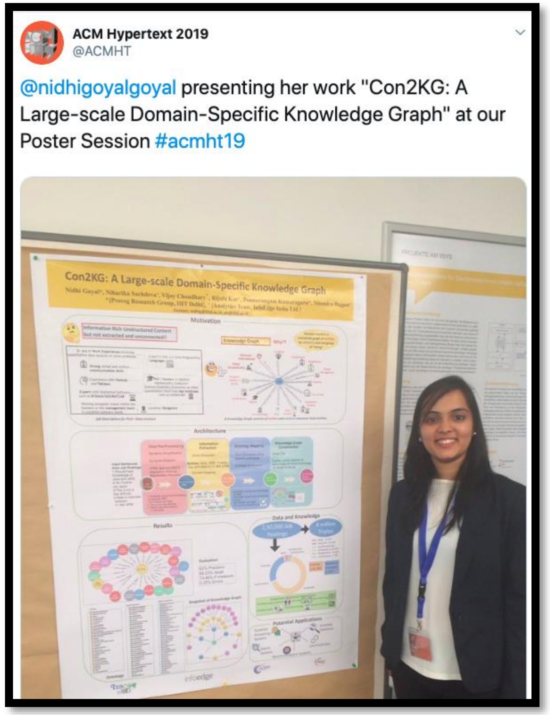
Publications

1. **Goyal, N.**, Sachdeva, N., Goel, A., Kalra, J., and Kumaraguru, P. KCNet: Kernel-based Canonicalization Network for entities in Recruitment Domain. In 30th International Conference on Artificial Neural Networks (ICANN). 2021.
2. **Goyal, N.**, Sachdeva, N., and Kumaraguru, P. Spy The Lie: Fraudulent Jobs Detection in Recruitment Domain using Knowledge Graphs. In 14th International Conference on Knowledge Science, Engineering and Management (KSEM 2021). 2021.
3. **Goyal, N.**, Sachdeva N., Choudhary V., Kar R., Kumaraguru P., and Rajput N. Con2KG-A Large-scale Domain-Specific Knowledge Graph. In Proceedings of the 30th ACM Conference on Hypertext and Social Media, pp. 287-288. 2019.
4. Arora, U.*, **Goyal, N.***, Goel, A., Sachdeva,N., Kumaraguru, P. Ask It Right! Identifying Low-Quality questions on Community Question Answering Services . In Proceedings of International Joint Conference on Neural Networks (IJCNN-2022), July 19 - July 23, Padua, Italy.

References

- [1] Noy, Natasha, et al. "Industry-scale Knowledge Graphs: Lessons and Challenges: Five diverse technology companies show how it's done." Queue 17.2 (2019): 48-75.
- [2] Wang, Ruijie, et al. "Acekg: A large-scale knowledge graph for academic data mining." Proceedings of the 27th ACM international conference on information and knowledge management. 2018.
- [3] Pan, Jeff Z., et al. "Content based fake news detection using knowledge graphs." *International semantic web conference*. Springer, Cham, 2018.
- [4] Vidros, Sokratis, et al. "Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset." *Future Internet* 9.1 (2017): 6.
- [5] Bhola, Akshay, et al. "Retrieving skills from job descriptions: A language model based extreme multi-label classification framework." *Proceedings of the 28th International Conference on Computational Linguistics*. 2020.
- [6] Liu, Liting, et al. "Hiring Now: A Skill-Aware Multi-Attention Model for Job Posting Generation." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.
- [7] Fatma, Nausheen, et al. "Canonicalizing knowledge bases for recruitment domain." *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, Cham, 2020.
- [8] Vashishth, Shikhar, Prince Jain, and Partha Talukdar. "Cesi: Canonicalizing open knowledge bases using embeddings and side information." *Proceedings of the 2018 World Wide Web Conference*. 2018.

Research Outcomes



30th ACM Hypertext Conference 2019, Hof, Germany

Received complimentary registration for travel award to attend [NIPS 2020]().

Mentor at [ACM Summer Workshop-IGDTUW](), 2020

RBCDSAI Web Science Symposium 2019, IIT Madras

Got selected in Fair Access Initiative to attend ACM Hypertext 2020.

Mentoring Ph.D. students in the Student Mentorship Program.

Acknowledgements

Thank you
for your attention!

Contribution 1:
Details about facts:

https://docs.google.com/presentation/d/1JPeZp1Kmj5BVku16XZR8gpo8xvxwpmi0PkWHq73DDoE/edit#slide=id.g54587baa50_0_14
Why KGs for fact checking?
Survey fact checking: https://arxiv.org/pdf/2002.00388.pdf
Survey knowledge graphs: https://arxiv.org/pdf/2002.00388.pdf

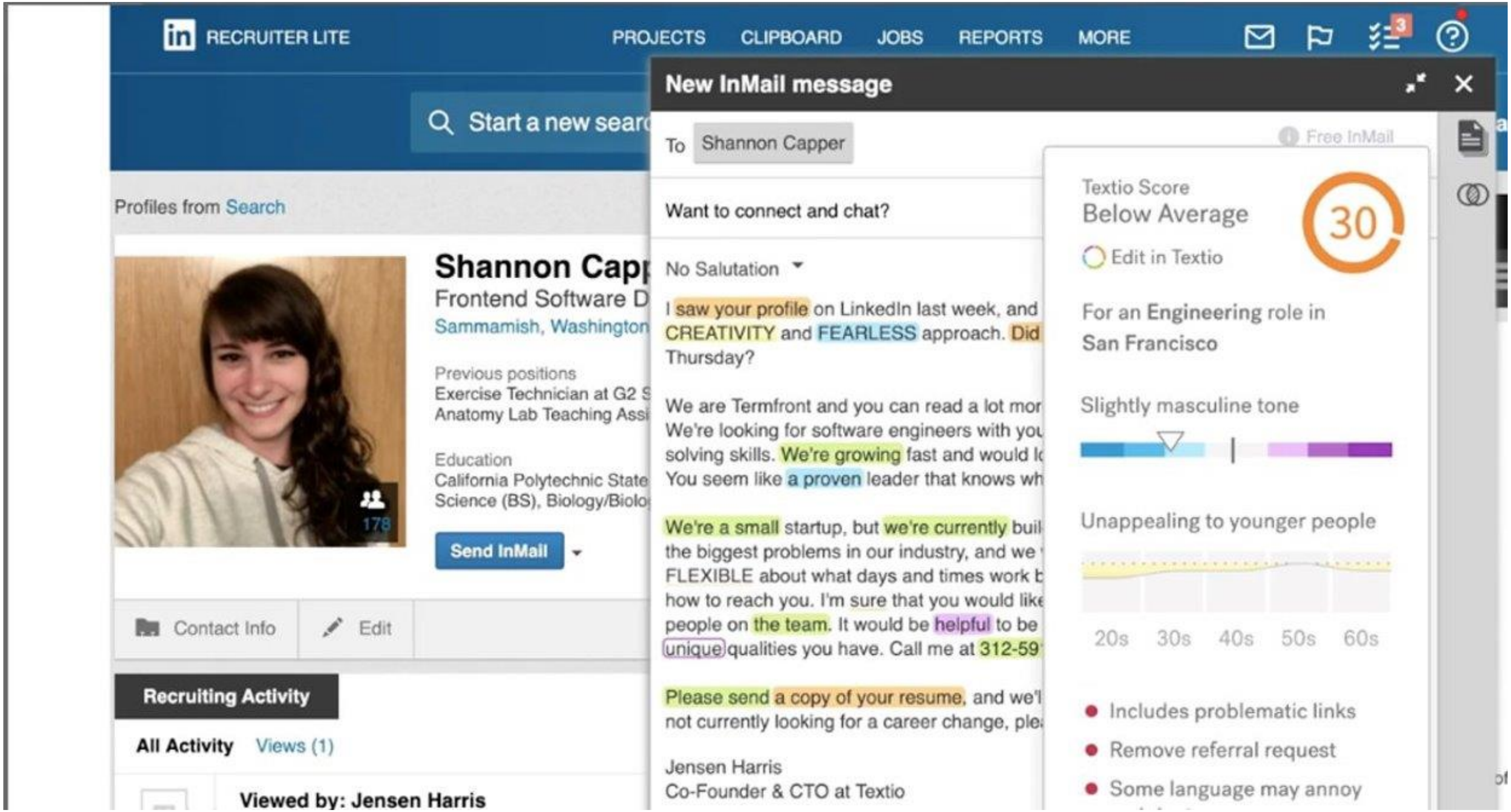
Contributions 2

- Slide 23:

Functions :

Table 3. Results of triple prediction task on proprietary dataset.

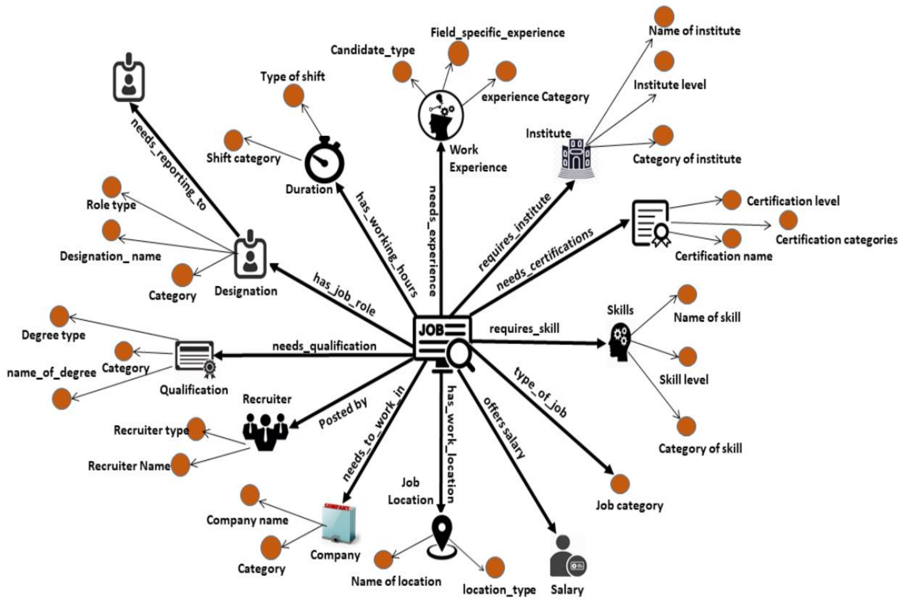
	MRR		Hits @		
Model	**Raw**	**Filter**	**1**	**3**	**10**
TransH	0.52	0.69	0.63	0.73	0.82
TransD	0.50	0.67	0.62	0.69	0.80
TransR	0.20	0.60	0.55	0.64	0.73
TransE	0.51	0.60	0.56	0.62	0.68
HolE	0.22	0.48	0.34	0.49	0.71
ComplEx	0.29	0.34	0.25	0.35	0.52
DisMult	0.30	0.40	0.30	0.40	0.50
RotatE	0.28	0.41	0.39	0.40	0.43


RECRUITER LITE
PROJECTS
CLIPBOARD
JOBS
REPORTS
MORE
Start a new searc
Profiles from Search
Shannon Capp
Frontend Software D
Sammamish, Washington
Previous positions
Exercise Technician at G2 S
Anatomy Lab Teaching Assi
Education
California Polytechnic State
Science (BS), Biology/Biolo
178
Send InMail
Contact Info
Edit
Recruiting Activity
All Activity
Views (1)
Viewed by: Jensen Harris
New InMail message
To
Shannon Capper
Free InMail
Want to connect and chat?
No Salutation
I saw your profile on LinkedIn last week, and
CREATIVITY and FEARLESS approach. Did
Thursday?
We are Termfront and you can read a lot mor
We're looking for software engineers with you
solving skills. We're growing fast and would l
You seem like a proven leader that knows wh
We're a small startup, but we're currently buil
the biggest problems in our industry, and we
FLEXIBLE about what days and times work b
how to reach you. I'm sure that you would like
people on the team. It would be helpful to be
unique qualities you have. Call me at 312-59
Please send a copy of your resume, and we'l
not currently looking for a career change, ple
Jensen Harris
Co-Founder & CTO at Textio
Textio Score
Below Average
30
Edit in Textio
For an Engineering role in
San Francisco
Slightly masculine tone
Unappealing to younger people
20s
30s
40s
50s
60s
Includes problematic links
Remove referral request
Some language may annoy

Entities	Count
Skills	12,057
Certifications	1100
Companies	2,23,955
Total Entities	3,65,061
Institutes	87,905
Designations	10,000
Qualifications	60
Total relations	40,11,030

Knowledge Graph

Graph structured knowledge bases (KBs) that store factual information in form of relationships between entities.

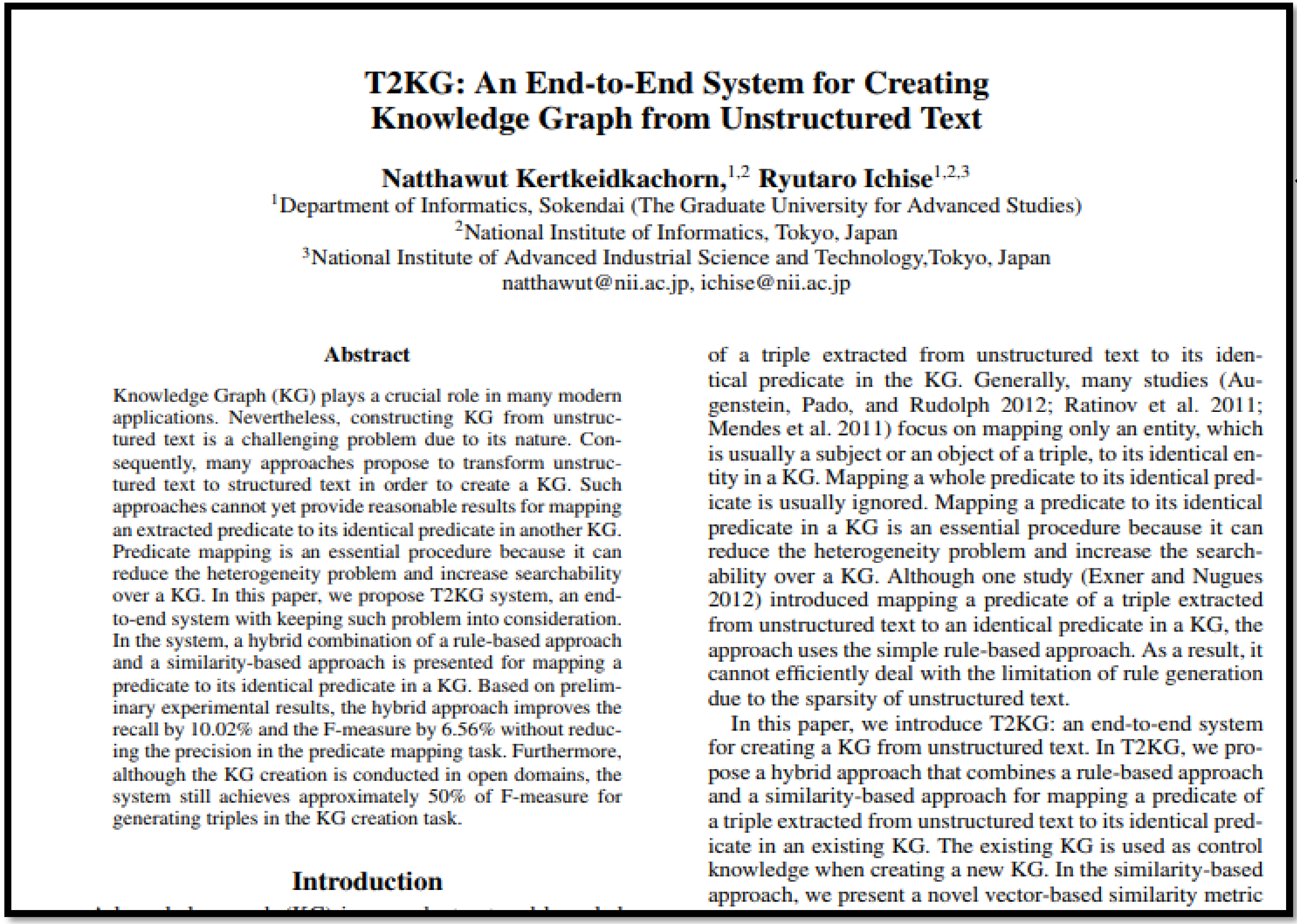


Nodes represent entities, edge labels represent types of relations, edges represent existing relationships.

Challenges

- Heterogeneous Data (different industries and business areas, languages, labour markets, educational systems etc.)
- Dynamically Evolving behavior of users
- Unavailability of Domain Specific Knowledge Bases
- Huge Volumes of Data- Recruitment Business with billions of users.

Literature Review



T2KG: An End-to-End System for Creating Knowledge Graph from Unstructured Text

Natthawut Kertkeidkachorn,[1,2] Ryutaro Ichise[1,2,3]
[1]Department of Informatics, Sokendai (The Graduate University for Advanced Studies)
[2]National Institute of Informatics, Tokyo, Japan
[3]National Institute of Advanced Industrial Science and Technology,Tokyo, Japan
natthawut@nii.ac.jp, ichise@nii.ac.jp

Abstract

Knowledge Graph (KG) plays a crucial role in many modern applications. Nevertheless, constructing KG from unstructured text is a challenging problem due to its nature. Consequently, many approaches propose to transform unstructured text to structured text in order to create a KG. Such approaches cannot yet provide reasonable results for mapping an extracted predicate to its identical predicate in another KG. Predicate mapping is an essential procedure because it can reduce the heterogeneity problem and increase searchability over a KG. In this paper, we propose T2KG system, an end-to-end system with keeping such problem into consideration. In the system, a hybrid combination of a rule-based approach and a similarity-based approach is presented for mapping a predicate to its identical predicate in a KG. Based on preliminary experimental results, the hybrid approach improves the recall by 10.02% and the F-measure by 6.56% without reducing the precision in the predicate mapping task. Furthermore, although the KG creation is conducted in open domains, the system still achieves approximately 50% of F-measure for generating triples in the KG creation task.

Introduction

of a triple extracted from unstructured text to its identical predicate in the KG. Generally, many studies (Augenstein, Pado, and Rudolph 2012; Ratinov et al. 2011; Mendes et al. 2011) focus on mapping only an entity, which is usually a subject or an object of a triple, to its identical entity in a KG. Mapping a whole predicate to its identical predicate is usually ignored. Mapping a predicate to its identical predicate in a KG is an essential procedure because it can reduce the heterogeneity problem and increase the searchability over a KG. Although one study (Exner and Nugues 2012) introduced mapping a predicate of a triple extracted from unstructured text to an identical predicate in a KG, the approach uses the simple rule-based approach. As a result, it cannot efficiently deal with the limitation of rule generation due to the sparsity of unstructured text.

In this paper, we introduce T2KG: an end-to-end system for creating a KG from unstructured text. In T2KG, we propose a hybrid approach that combines a rule-based approach and a similarity-based approach for mapping a predicate of a triple extracted from unstructured text to its identical predicate in an existing KG. The existing KG is used as control knowledge when creating a new KG. In the similarity-based approach, we present a novel vector-based similarity metric

- Proposed an end-to-end framework for Information Extraction.
- Addressed the problem of predicate mapping that will reduce heterogeneity in KGs .
- Dataset: 1,20,000 Wikipedia articles
- Precision, Recall improved- **0.24** , **10.02**
- F- measure improved - **6.56**

Literature Review

AceKG: A Large-scale Knowledge Graph for Academic Data Mining

Ruijie Wang, Yuchen Yan, Jialu Wang, Yuting Jia, Ye Zhang, Weinan Zhang, Xinbing Wang
Shanghai Jiao Tong University, Shanghai, China
200240
{wjerry5,wnzhang,xwang8}@sjtu.edu.cn

ABSTRACT

Most existing knowledge graphs (KGs) in academic domains suffer from problems of insufficient multi-relational information, name ambiguity and improper data format for large-scale machine processing. In this paper, we present AceKG, a new large-scale KG in academic domain. AceKG not only provides clean academic information, but also offers a large-scale benchmark dataset for researchers to conduct challenging data mining projects including link prediction, community detection and scholar classification. Specifically, AceKG describes 3.13 billion triples of academic facts based on a consistent ontology, including necessary properties of papers, authors, fields of study, venues and institutes, as well as the relations among them. To enrich the proposed knowledge graph, we also perform entity alignment with existing databases and rule-based inference. Based on AceKG, we conduct experiments of three typical academic data mining tasks and evaluate several state-of-the-art knowledge embedding and network representation learning approaches on the benchmark datasets built from AceKG. Finally, we discuss several promising research directions that benefit from AceKG.

KEYWORDS

Knowledge Graphs, Academic Data Mining, Benchmarking

aim at discovering cross-field knowledge [12]. Third, synonymy and ambiguity are also the restrictions for knowledge mining [13]. Allocating the unique IDs to the entities is the necessary solution, but some databases use the names of the entities as their IDs directly.

In this paper, we propose Academic Knowledge Graph (AceKG), [1] an academic semantic network, which describes 3.13 billion triples of academic facts based on a consistent ontology, including commonly used properties of papers, authors, fields of study, venues, institutes and relations among them. Apart from the knowledge graph itself, we also perform entity alignment with the existing KGs or datasets and some rule-based inferences to further extend it and make it linked with other KGs in the linked open data cloud. Based on AceKG, we further evaluate several state-of-the-art knowledge embedding and network representation learning approaches in Sections 3 and 4. Finally we discuss several potential research directions that benefit from AceKG in Section 5 and conclude in Section 6.

Compared with other existing open academic KGs or datasets, AceKG has the following advantages.

(1) AceKG offers a heterogeneous academic information network, i.e., with multiple entity categories and relationship types, which supports researchers or engineers to conduct various academic data mining experiments.
(2) AceKG is sufficiently large (3.13 billion triples with nearly 100G disk size) to cover most instances in the academic ontology,

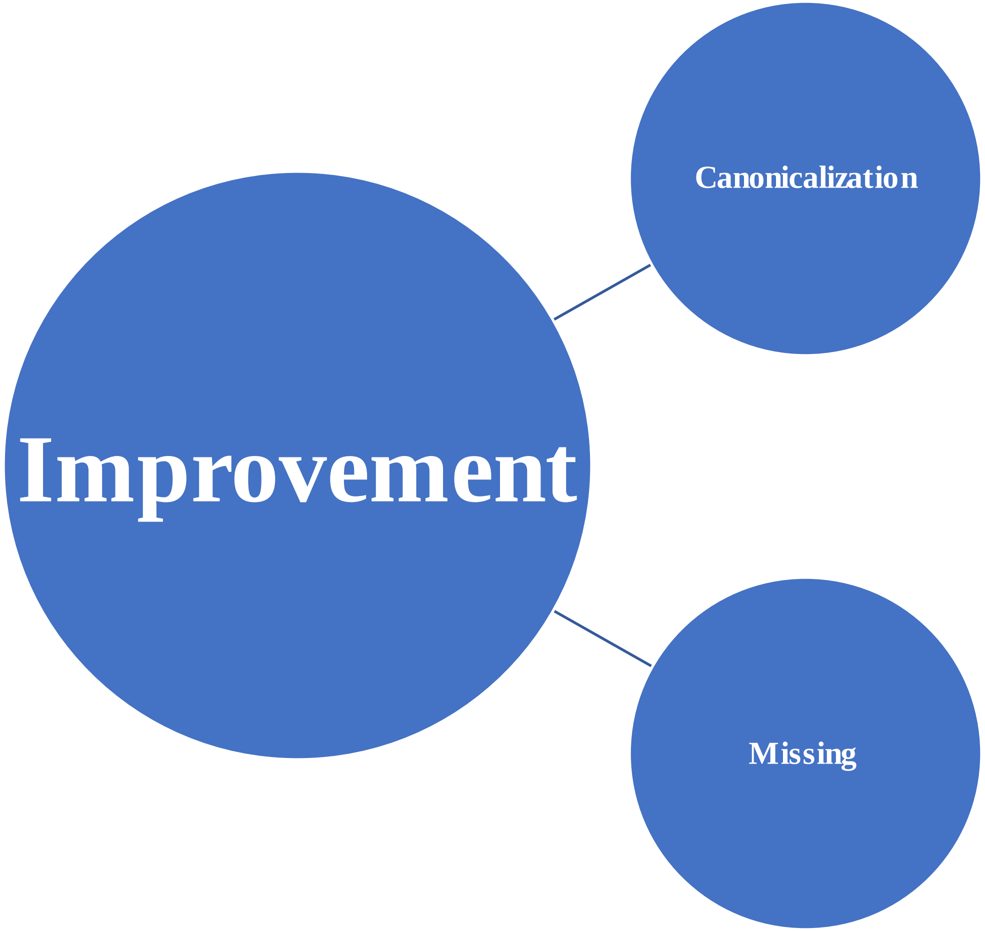
- Heterogeneous Academic Information Network.
- Dataset: **3.13 billion triples**.
- Extracted all scholars, papers and venues in those fields of study to construct 5 heterogeneous collaboration networks.

Side Information Collection

- **We acquired additional knowledge using:**
- **Wikipedia InfoBox:** Extracted knowledge from Wikipedia infoboxes for different datasets.
- {‘title wikis’, ‘websites’, ‘types’} - RDE(S) {‘Names’, ‘websites’, ‘title wikis’} - RDE(D) {‘Names’, ‘websites’, ‘affiliation’} - RDE(I) {‘Names’, ‘websites’, ‘title wikis’, ‘types’} - ESCO(S) {‘Names’, ‘websites’, ‘title wikis’} - ESCO(D) {‘types’, ‘industries’, ‘websites’, ‘native names’, ‘title wikis’} - DBpedia(C).

- **Google Knowledge graph (Serp API):** We extract textual descriptions and other attributes such as {location, type, established} for entities to supplement the model with semantic knowledge.

Literature Review



- Generic approaches → Kertkeidkachorn et al. (AAAI 2017)
- Domain-specific methods → Wang et al. (CIKM 2018)

- Content-based approaches → Vidros et al. (Future Internet 2017)
- Context-based approaches → Mahbub et al. (ISD 2018)
- Graph-based approaches → Pan et al. (ISWC 2018)

Contribution
3: Improve quality of job postings

- **We acquired additional knowledge using:**
- **Wikipedia InfoBox:** Extracted knowledge from Wikipedia infoboxes for different datasets.
- {‘title wikis’, ‘websites’, ‘types’} - RDE(S) {‘Names’, ‘websites’, ‘title wikis’} - RDE(D) {‘Names’, ‘websites’, ‘affiliation’} - RDE(I) {‘Names’, ‘websites’, ‘title wikis’, ‘types’} - ESCO(S) {‘Names’, ‘websites’, ‘title wikis’} - ESCO(D) {‘types’, ‘industries’, ‘websites’, ‘native names’, ‘title wikis’} - DBpedia(C).

- **Google Knowledge graph (Serp API):** We extract textual descriptions and other

What to Identify?

- Fraudulent jobs are dishonest, money seeking, intentionally and verifiably false that mislead job seekers.
- Fraudulent jobs contain untenable facts about domain-specific entities such as mismatch in skills, industries, offered compensation, etc.

Data Entry Clerks Position
We have several openings available in this area earning $1000.00-$2500.00 per week. We are seeking only honest, self-motivated people with a desire to work in the home typing and data entry field, from the comfort of their own homes.The preferred applicants should be at least 18 years old with Internet access. No experience is needed. However the following skills are desirable: Basic computer and typing skills, ability to spell and print neatly, ability to follow directions.
Earn as much as you can from the comfort of your home typing and doing data entry.
You do NOT need any special skills to get started.

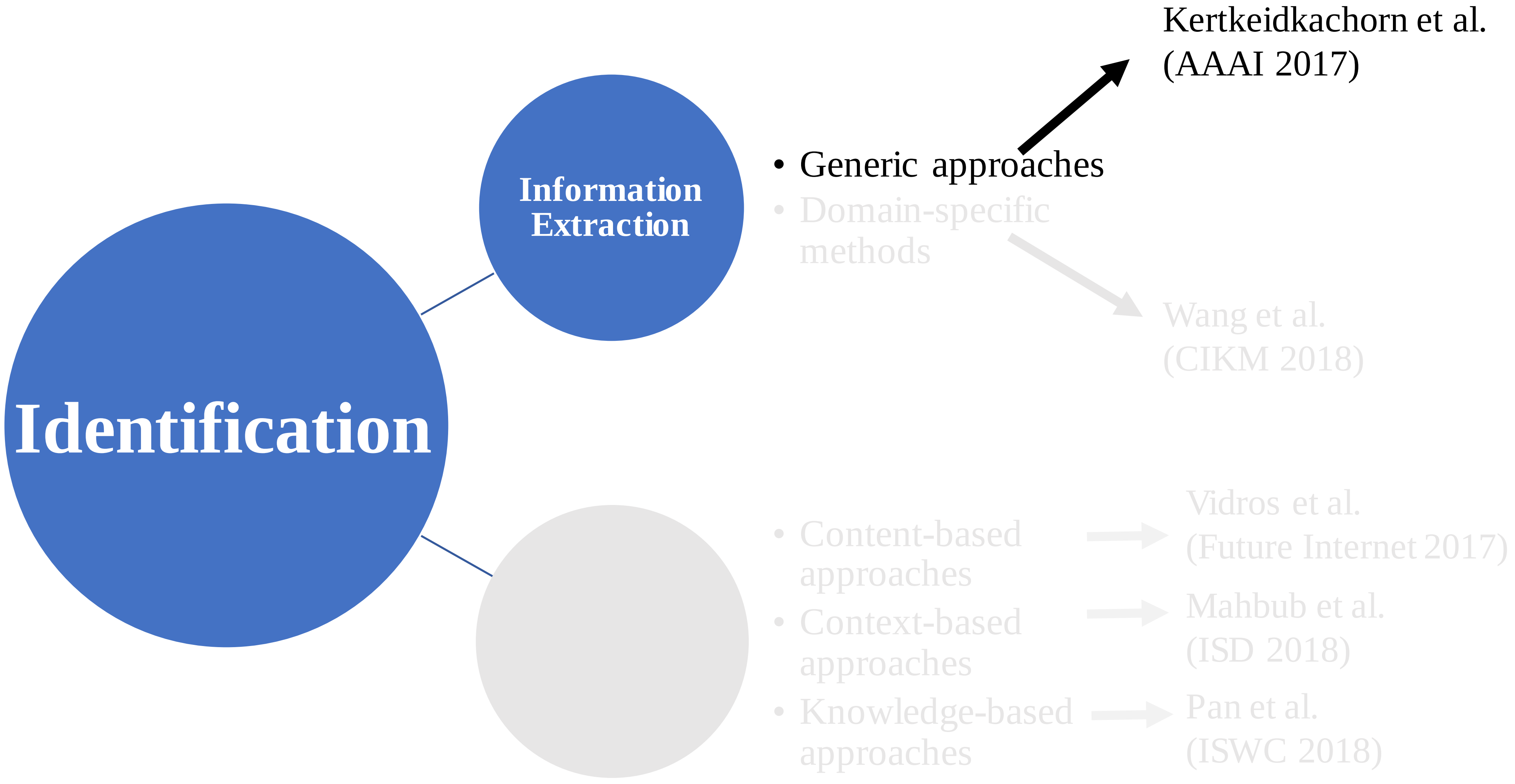
Data Entry Clerk
Responsibilities include, but are not limited to:
Review and process confidential and extremely time-sensitive applications.
Identify objective data and enter (""key what you see"") at a high level of productivity and accuracy.
Perform data entry task from a paper and/or document image.
Utilize system functions to perform data look-up and validation.
High volume sorting, analyzing, indexing, of insurance, legal and financial documents.
Maintain high degree of quality control and validation of the completed work
Identify, classify, and sort documents electronically.

Fig. 1. Examples of job postings a) fraudulent job on the left and b) legitimate at the right. These job postings are taken from publicly available dataset.

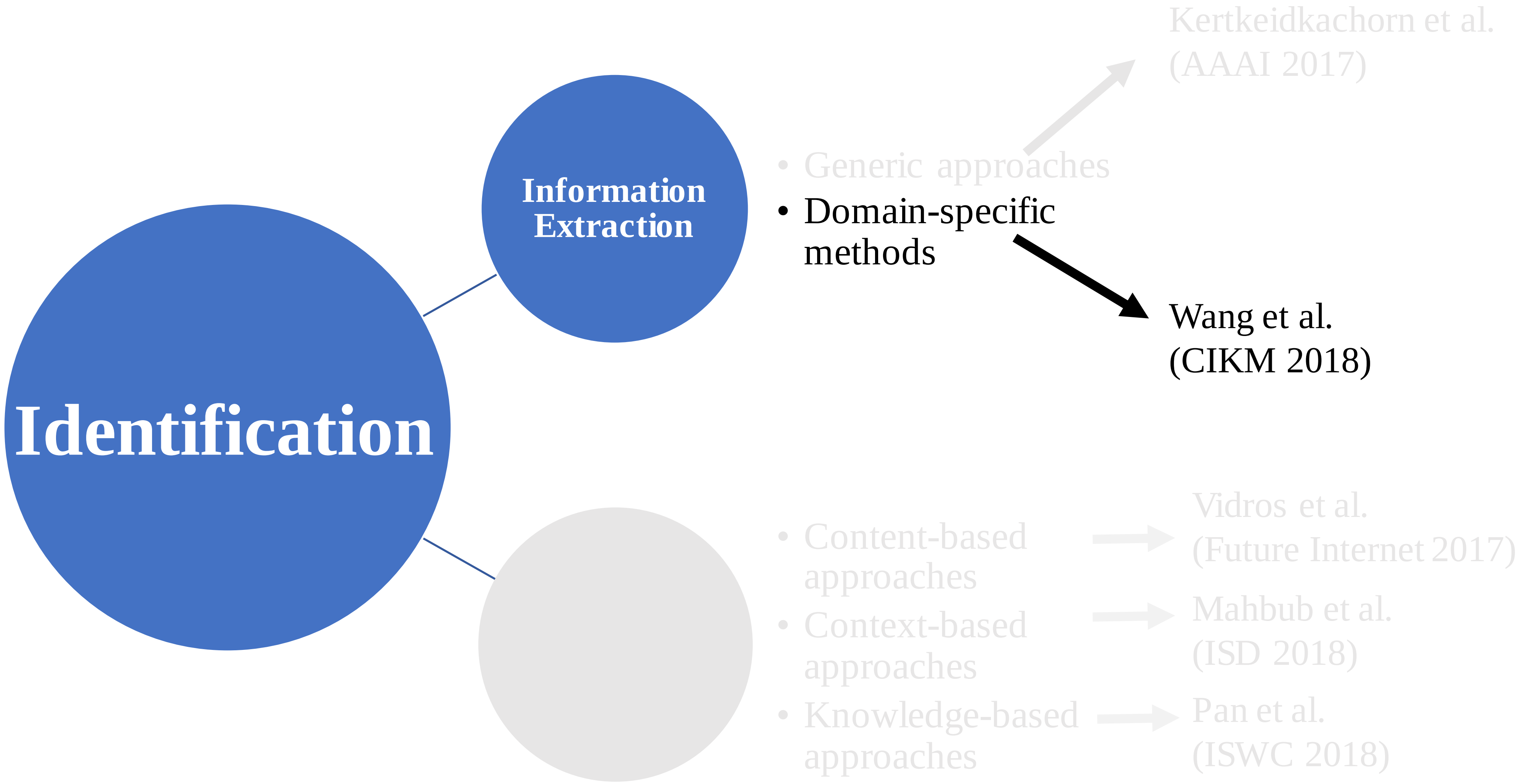
Literature Review

<table>
<tr><th>Paper title/ Reference</th><th>Domain / Criteria</th><th>Research gap</th></tr>
<tr><td>CESI: Canonicalizing open knowledge bases using embeddings and side information [8]
(WWW, 2018)</td><td rowspan="2">Non-standard</td><td rowspan="2">Recent research discusses either statistical similarity measures or deep learning methods like word-embedding or siamese network-based representations for canonicalization.</td></tr>
<tr><td>Canonicalization of entities in recruitment domain [7]
(PAKDD, 2020)</td></tr>
<tr><td>Hiring Now A Skill-Aware Multi-Attention Model for Job Posting Generation [6]
(ACL, 2020)</td><td rowspan="2">Missing</td><td rowspan="2">Existing approaches are limited to contextual modelling and do not exploit inter-relational structures such as job-job and job-skill relationships.</td></tr>
<tr><td>Retrieving Skills from Job Descriptions: A Language Model Based Extreme Multi-label Classification Framework [5]
(COLING, 2020)</td></tr>
</table>

Literature Review



Literature Review



Research Objectives

1. To Identify misleading content
    - Extract domain-specific information from job postings and construct domain-specific knowledge base.
    - Build a framework to classify misleading information using domain knowledge.
2. To Improve job posting quality
    - Standardize the recruitment domain entities (skills, institutes, companies, designations).
    - Build a framework for missing entities (skills) prediction.

Research Objectives

1. To Identify misleading content
   - Extract domain-specific information from job postings and construct domain-specific knowledge base.
   - Build a framework to classify misleading content using domain knowledge.
2. To Improve job posting quality
   - Standardize the recruitment domain entities (skills, institutes, companies, designations).
   - Build a framework for missing entities (skills) prediction.

Literature Review

Identification

- Generic approaches
- Domain-specific methods

Kertkeidkachorn et al. (AAAI 2017)

Wang et al. (CIKM 2018)

Misleading Content

- Content-based approaches
- Context-based approaches
- Knowledge-based approaches

Vidros et al. (Future Internet 2017)

Mahbub et al. (ISD 2018)

Pan et al. (ISWC 2018)

Literature Review

Identi

Content Based Fake News Detection Using Knowledge Graphs

Jeff Z. Pan[1(✉)], Siyana Pavlova[1], Chenxi Li[1,2], Ningxi Li[1,2], Yangmei Li[1,2], and Jinshuo Liu[2(✉)]

[1] University of Aberdeen, Aberdeen, UK
jeff.z.pan@abdn.ac.uk
[2] Wuhan University, Wuhan, China
liujinshuo@whu.edu.cn

Abstract. This paper addresses the problem of fake news detection. There are many works already in this space; however, most of them are for social media and not using news content for the decision making. In this paper, we propose some novel approaches, including the B-TransE model, to detecting fake news based on news content using knowledge graphs. In our solutions, we need to address a few technical challenges. Firstly, computational-oriented fact checking is not comprehensive enough to cover all the relations needed for fake news detection. Secondly, it is challenging to validate the correctness of the extracted triples from news articles. Our approaches are evaluated with the Kaggle's 'Getting Real about Fake News' dataset and some true articles from main stream media. The evaluations show that some of our approaches have over 0.80 F1-scores.

- Fake news Detection Problem.
- Proposed B-TransE model to detect fake news using knowledge graphs.
- Addressed the problem of computational-oriented fact checking.
- Dataset: Kaggle "Getting real about fake news".
- F- measure improved – **0.81**

Literature Review

Identification

Kertkeidkachorn et al. (AAAI 2017)

- Generic approaches
- Domain-specific methods

Wang et al. (CIKM 2018)

Misleading Content

- Content-based approaches
- Context-based approaches
- Knowledge-based approaches

Vidros et al. (Future Internet 2017)

Mahbub et al. (ISD 2018)

Pan et al. (ISWC 2018)

Research Objectives

1. To Identify misleading content
    - Extract domain-specific information from job postings and construct domain-specific knowledge base.
    - Build a framework to classify misleading content using domain knowledge.
2. To Improve job posting quality
    - Standardize the recruitment domain entities (skills, institutes, companies, designations).
    - Build a framework for missing entities (skills) prediction.

Literature Review

Identi

Content Based Fake News Detection Using Knowledge Graphs

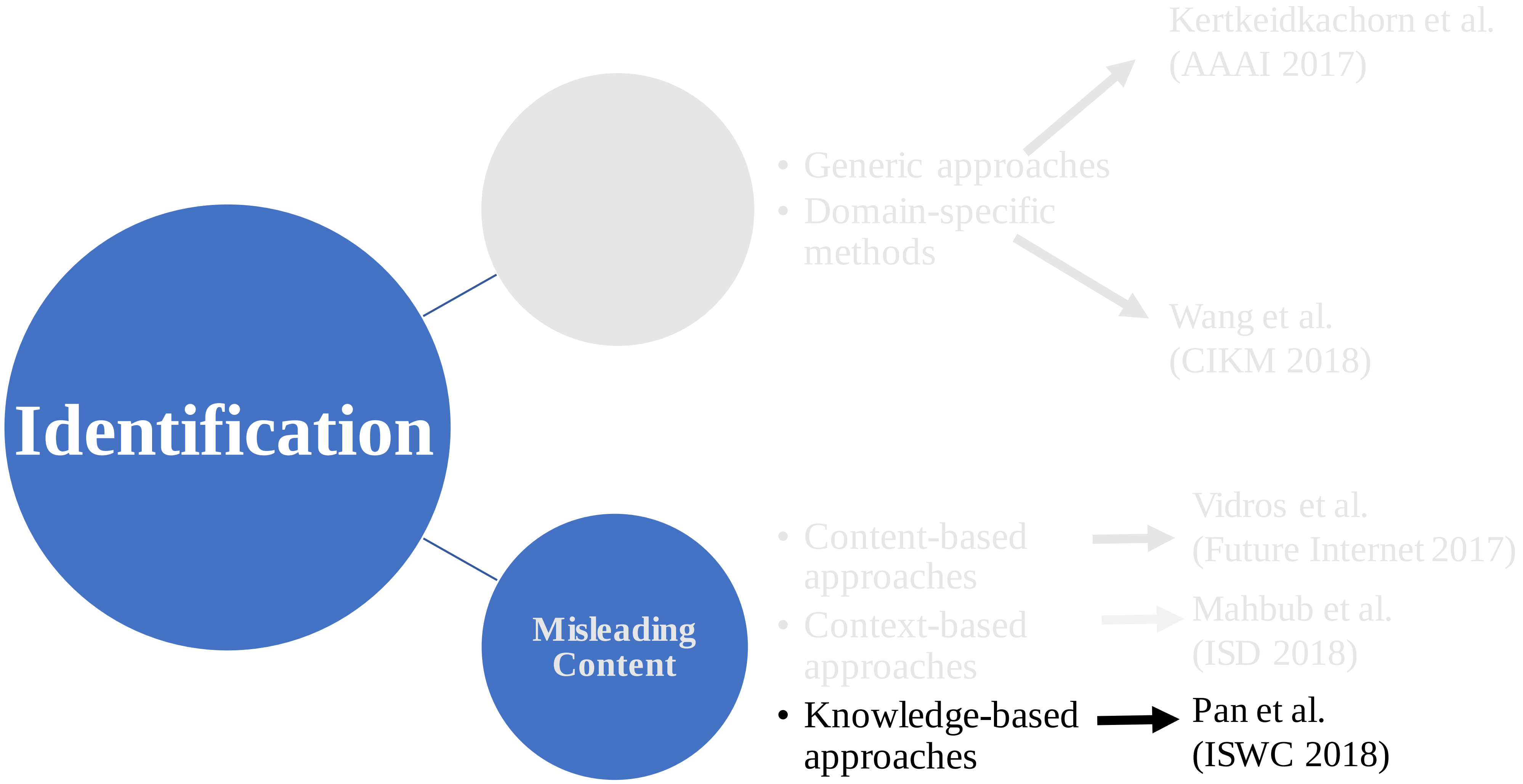
Jeff Z. Pan[1(✉)], Siyana Pavlova[1], Chenxi Li[1,2], Ningxi Li[1,2], Yangmei Li[1,2], and Jinshuo Liu[2(✉)]

[1] University of Aberdeen, Aberdeen, UK
jeff.z.pan@abdn.ac.uk
[2] Wuhan University, Wuhan, China
liujinshuo@whu.edu.cn

Abstract. This paper addresses the problem of fake news detection. There are many works already in this space; however, most of them are for social media and not using news content for the decision making. In this paper, we propose some novel approaches, including the B-TransE model, to detecting fake news based on news content using knowledge graphs. In our solutions, we need to address a few technical challenges. Firstly, computational-oriented fact checking is not comprehensive enough to cover all the relations needed for fake news detection. Secondly, it is challenging to validate the correctness of the extracted triples from news articles. Our approaches are evaluated with the Kaggle's 'Getting Real about Fake News' dataset and some true articles from main stream media. The evaluations show that some of our approaches have over 0.80 F1-scores.

- Fake news Detection Problem.
- Proposed B-TransE model to detect fake news using knowledge graphs.
- Addressed the problem of computational-oriented fact checking.
- Dataset: Kaggle "Getting real about fake news".
- F- measure improved – **0.81**

Literature Review



Literature Review

Identi

Content Based Fake News Detection Using Knowledge Graphs

Jeff Z. Pan[1(✉)], Siyana Pavlova[1], Chenxi Li[1,2], Ningxi Li[1,2], Yangmei Li[1,2], and Jinshuo Liu[2(✉)]

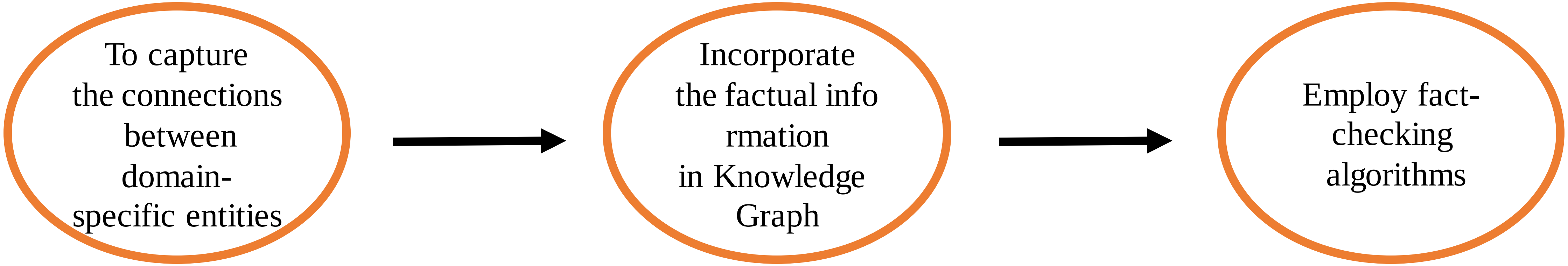
[1] University of Aberdeen, Aberdeen, UK
jeff.z.pan@abdn.ac.uk
[2] Wuhan University, Wuhan, China
liujinshuo@whu.edu.cn

Abstract. This paper addresses the problem of fake news detection. There are many works already in this space; however, most of them are for social media and not using news content for the decision making. In this paper, we propose some novel approaches, including the B-TransE model, to detecting fake news based on news content using knowledge graphs. In our solutions, we need to address a few technical challenges. Firstly, computational-oriented fact checking is not comprehensive enough to cover all the relations needed for fake news detection. Secondly, it is challenging to validate the correctness of the extracted triples from news articles. Our approaches are evaluated with the Kaggle's 'Getting Real about Fake News' dataset and some true articles from main stream media. The evaluations show that some of our approaches have over 0.80 F1-scores.

- Fake news Detection Problem.
- Proposed B-TransE model to detect fake news using knowledge graphs.
- Addressed the problem of computational-oriented fact checking.
- Dataset: Kaggle "Getting real about fake news".
- F- measure improved – **0.81**

Contribution 2

- Existing approaches mainly focus on handcrafted, linguistic, writing styles, string-based features of job postings.
- Ignore the factual information among domain-specific entities present in job postings, which are important to capture relationships.



Related Work

Related work	Domain / Criteria	Research gap
Kertkeidkachorn et al. , T2KG: An End-to-End System for Creating Knowledge Graph (AAAI, 2017)	Domain-specific Knowledge Graphs	1. Open (Public) Knowledge bases are available. They do not contain domain-specific information. 2. Recruitment domain-specific Knowledge bases are unavailable.
Wang et al. (AceKG: A large-scale Knowledge Graph (CIKM, 2018)		
Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset [4] (Future Internet, 2017)	Misleading	1. Existing approaches focus on studying writing styles, linguistics, and context-based features. 2. Ignore the relationships among domain-specific entities. 3. Unavailability of recruitment domain Knowledge Graph.
Content-based fake news Detection [3] (ISWC, 2020)		

- In future,
  - Plan to test our approach for hierarchy-based, neural network-based and path-based fact-checking algorithms.
  - Learning heterogeneous information from documents such as CVs to build an integrated framework and explore user features.

Research Work

1. To Identify misleading content
    - Extract domain-specific information from job postings and construct domain-specific knowledge base.
    - Build a framework to classify misleading information using domain knowledge.
2. To Improve job posting quality
    - Standardize the recruitment domain entities (skills, institutes, companies, designations).
    - Build a framework for missing entities (skills) prediction.

Problem Formulation

Let J = {J_1, J_2, J_3........., J_N} be the set of job postings and Y = {y_1, y_2, y_3........., y_n} be corresponding labels such that $y_i \in \{0, 1\}$. For every J_i, we extracted a set of triples T^i where $T^i = \{t^i_1, t^i_2, t^i_3, \ldots\ldots\ldots, t^i_k\}$ and $k > 0$; using OpenIE. A triple $t^i_j \in T^i$ is of the form (subject (s), predicate (p), object (o)) where (s, o) $\in$ E and $p \in$ P. We further define $m^i \in M$ and $c^i \in C$ as meta features and contextual features extracted from J_i

Summary

- We design a novel multi-tier framework Kernel-based Canonicalization Network (KCNet).
- KCNet induces a non-linear mapping between the contextual vector representations while capturing fine-granular and high-dimensional relationships among vectors.
- KCNet efficiently models more prosperous semantic and meta side information from external knowledge towards exploring kernel features for canonicalizing entities in the recruitment domain.
- We demonstrate that our proposed methods are also generalizable to domain-specific entities in similar scenarios.

Objective

Our objective is to learn function Φ where Φ: F ($KG^{A}_{false}(T)^{i}$, $KG^{A}_{true}(T)^{i}$, c^{i}, m^{i} where $KG^{A}_{true}(T)^{i}$ is the scoring function, we learn from triple $t^{i} \in T^{i} | y_i = 0$ of legitimate job postings and $KG^{A}_{false}(T)^{i}$ from triple $t^{i} \in T^{i} | y_i = 1$ of fraudulent job postings. Here $KG^{A} \in$ {*TransE*, *TransR*, *TransH*, *TransD*, *DistMult*, *ComplEx*, *HolE*, *RotatE*} which are popular fact-checking algorithms from existing knowledge graph literature.

- https://precog.iiitd.edu.in/pubs/2021_July_KCNet.pdf