

Improving *Content Quality* for *Online Professional Activities* using *Domain Specific Learning and Knowledge*



PhD Thesis Defense By

Nidhi Goyal

Advisors: Prof. Ponnurangam Kumaraguru, Dr. V. Raghava Mutharaju, Dr. Niharika Sachdeva



Evaluation Committee



Prof. Mehwish Alam
Télécom Paris
Institut Polytechnique de Paris



Prof. Amit Awekar
IIT Guwahati



Prof. Maya Ramanath
IIT Delhi

Thesis Supervisors



Prof. Ponnuram Kumaraguru ("PK")
IIIT-Hyderabad



Dr. V. Raghava Mutharaju
IIIT-Delhi



Dr. Niharika Sachdeva
InfoEdge India Limited

Outline

- 1 Online Professional Activities
- 2 Content Quality Issues
- 3 Thesis Question
- 4 Thesis Contributions
- 5 Proposed methodologies
- 6 Conclusion

Online Professional Activities



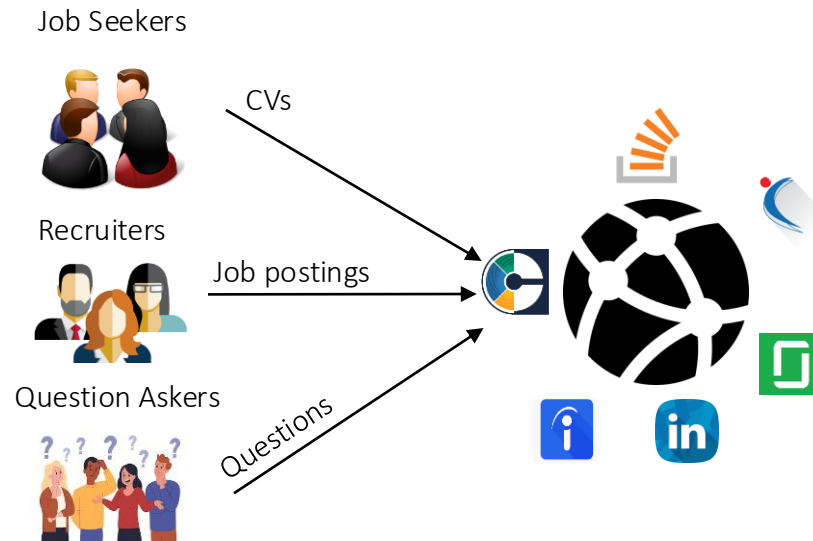
Online Professional Activities

Massive increase in professional, scientific, and technical services from 2021-2031



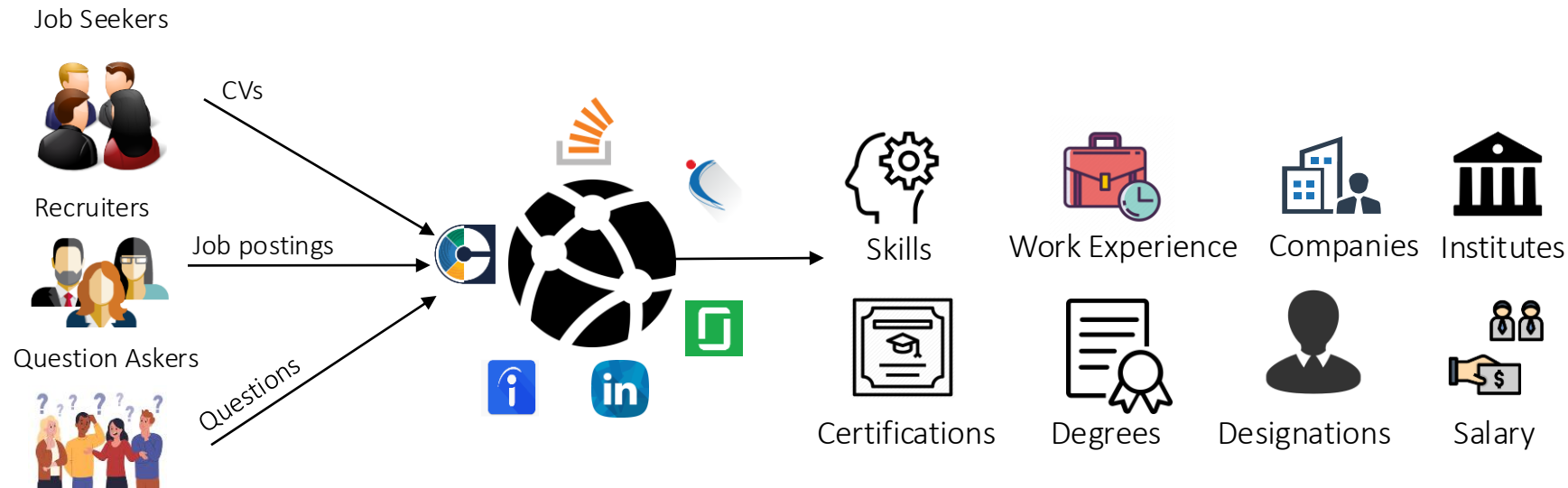
Online Professional Activities

Massive increase in professional, scientific, and technical services from 2021-2031



Online Professional Activities

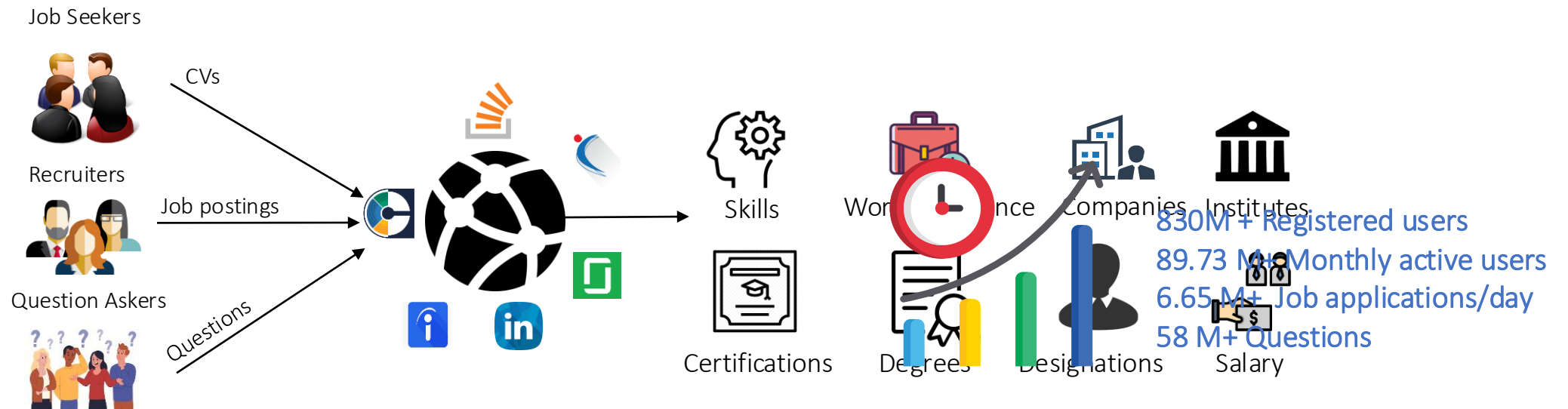
Massive increase in professional, scientific, and technical services from 2021-2031



Interaction of different contributors and the content shared by them on online professional platforms

Online Professional Activities

Massive increase in professional, scientific, and technical services from 2021-2031



Interaction of different contributors and the content shared by them on online professional platforms

Content Quality Matters!



830 M+ Registered users
89.73 M+ Monthly active users
6.65 M+ Job applications/day
58 M+ Questions



Content Quality Matters!

Quality follows the value content create after being put to use
Content Quality is paramount for those who rely on online professional platforms for business & networking

How often should I post on LinkedIn to stay active?



There's no one-size-fits-all answer, but aiming for 1-2 posts per week can help you stay visible without overwhelming your network. Quality over quantity is important; focus on sharing insightful content that adds value to your audience.

Content Quality Issues



Content Quality Issues

1 Inconsistent Content (Non-standard Entities)



Dr. Babasaheb Ambedkar
Marathwada University
Aurangabad
1145 variations



ICICI Prudential Life Insurance
497 variations

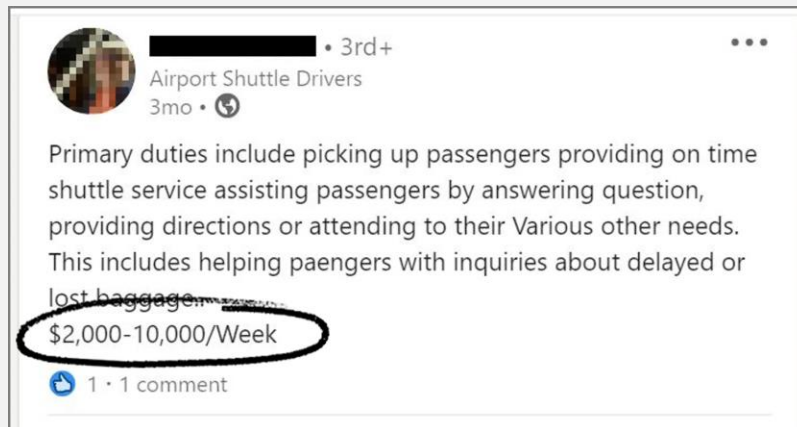
2 Incomplete Content (Missing Entities)

Job Title	Market Analyst
Job description	Assist the Manager in sourcing food industry, in conducting product research and analysis. Facilitate effective communication between the analytics and user experience teams. Strong research, data analysis and communication skills.
Required skills	communication data analysis regex visualization python
<div><div>Explicit Skills</div><div>Implicit Skills</div></div>	

Around 65% of job descriptions are missing skills

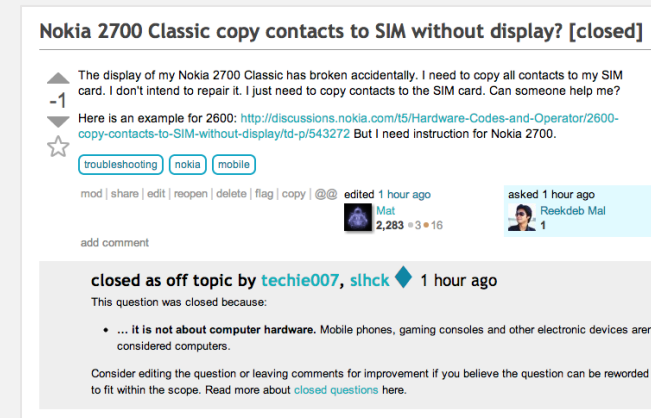
Content Quality Issues

3 Fraudulent Content



A misleading job advertisement on LinkedIn that promises unusually high salary for shuttle-bus drivers

4 Low-quality Content



A question gets closed due to off-topic content on technical question answering website

Core Thesis Question & Thesis Contributions

Core Thesis Question

How can we **improve** the quality of online professional content by leveraging domain-specific learning and knowledge?

Thesis Contributions

1

Developed a multi-tier framework, KCNet to normalize domain-specific entities (skills, institutes, companies, and designations)

2

Proposed and evaluated novel framework, JobXMLC for finding missing entities to improve the quality of jobs using job-skill graph

3

Built a Domain-specific Knowledge Graph (Con2KG) and developed a novel framework FRJD to classify fraudulent and legitimate jobs

4

Developed an architecture to identify low quality (off-topic, too broad, opinion-based, unclear what are you asking) questions

Thesis Contributions

1

Developed a multi-tier framework, KCNet to normalize domain-specific entities (skills, institutes, companies, and designations)

Inconsistent Content

Non-standard Entities

01. Spelling Variations	<ul style="list-style-type: none"> • Java Developer • Java Deveoper
02. Hierarchical variations	<ul style="list-style-type: none"> • Oracle Financial Services Software • Oracle Corporation
03. Overlapping but different entities	<ul style="list-style-type: none"> • Emerald Bikes pvt limited • Emerald Jewellery Retail Limited
04. Domain specific concepts	<ul style="list-style-type: none"> • SOAP • REST
05. Semantic variations	<ul style="list-style-type: none"> • Accel Frontline • Inspirisys
06. Short Forms or Abbreviations	<ul style="list-style-type: none"> • umbc • University of Maryland, Baltimore County

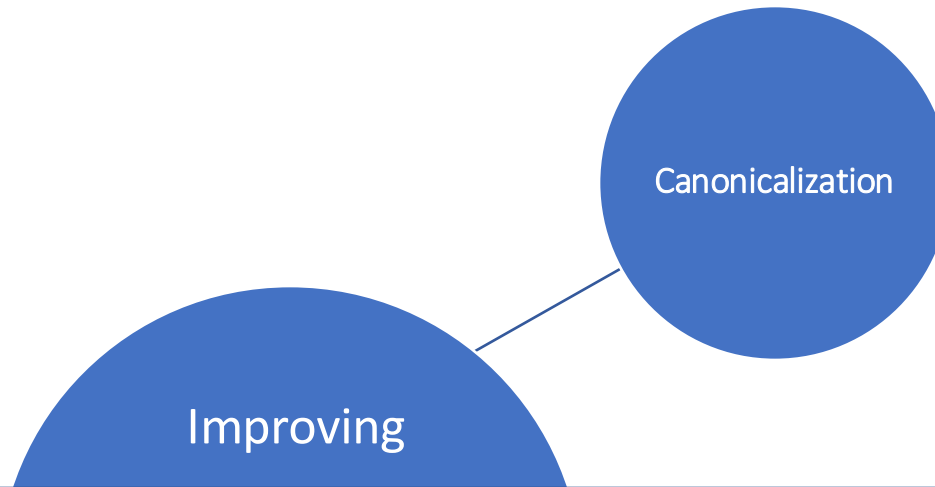
Objective

unive
Umb

Microsoft corp india pvt ltd
Microsoft corporation (india) pvt. ltd
Microsoft corp india pvt ltd
Microsoft india



Research Gap



- Generic approaches
- Domain-specific methods

Vashishtha et al.
(WWW 2018)

Fatma et al.
(PAKDD 2020)

Focus upon either statistical similarity measures or deep learning methods like word-embedding and lack domain-specific knowledge for normalization.

Dataset Statistics

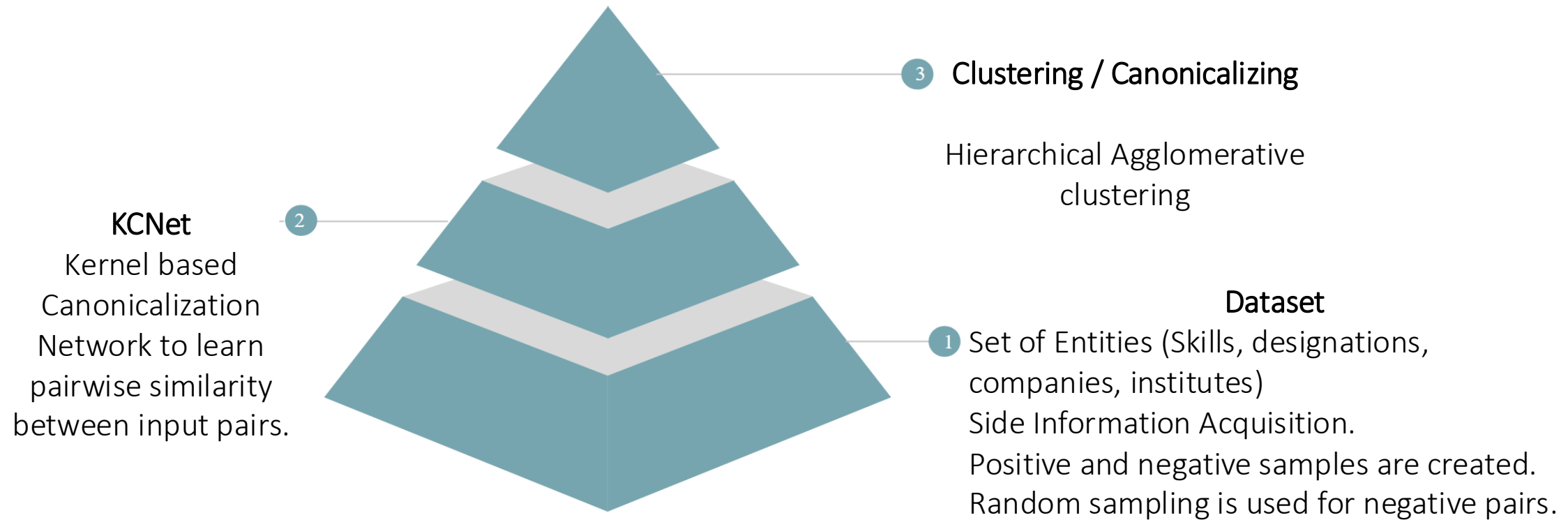
Source	Dataset	Entity Clusters
Proprietary	RDE(C)	25,602
	RDE(I)	23,690
	RDE(D)	3,894
	RDE(S)	607
Open	DBpedia(C)	2,944
	ESCO(S)	2,644
	ESCO(D)	2,903

Dataset from popular recruitment platform

Source	Dataset	Side Information
Proprietary	RDE(C)	{‘title wikis’, ‘websites’}
	RDE(I)	{‘Names’, ‘websites’, ‘affiliation’}
	RDE(D)	{‘Names’, ‘websites’, ‘title wikis’}
	RDE(S)	{‘title wikis’, ‘websites’, ‘types’}
Open	DBpedia (C)	{‘types’, ‘industries’, ‘websites’, ‘native names’, ‘title wikis’}
	ESCO(S)	{‘Names’, ‘title wikis’, ‘websites’, ‘types’}
	ESCO(D)	{‘Names’, ‘websites’, ‘title wikis’}

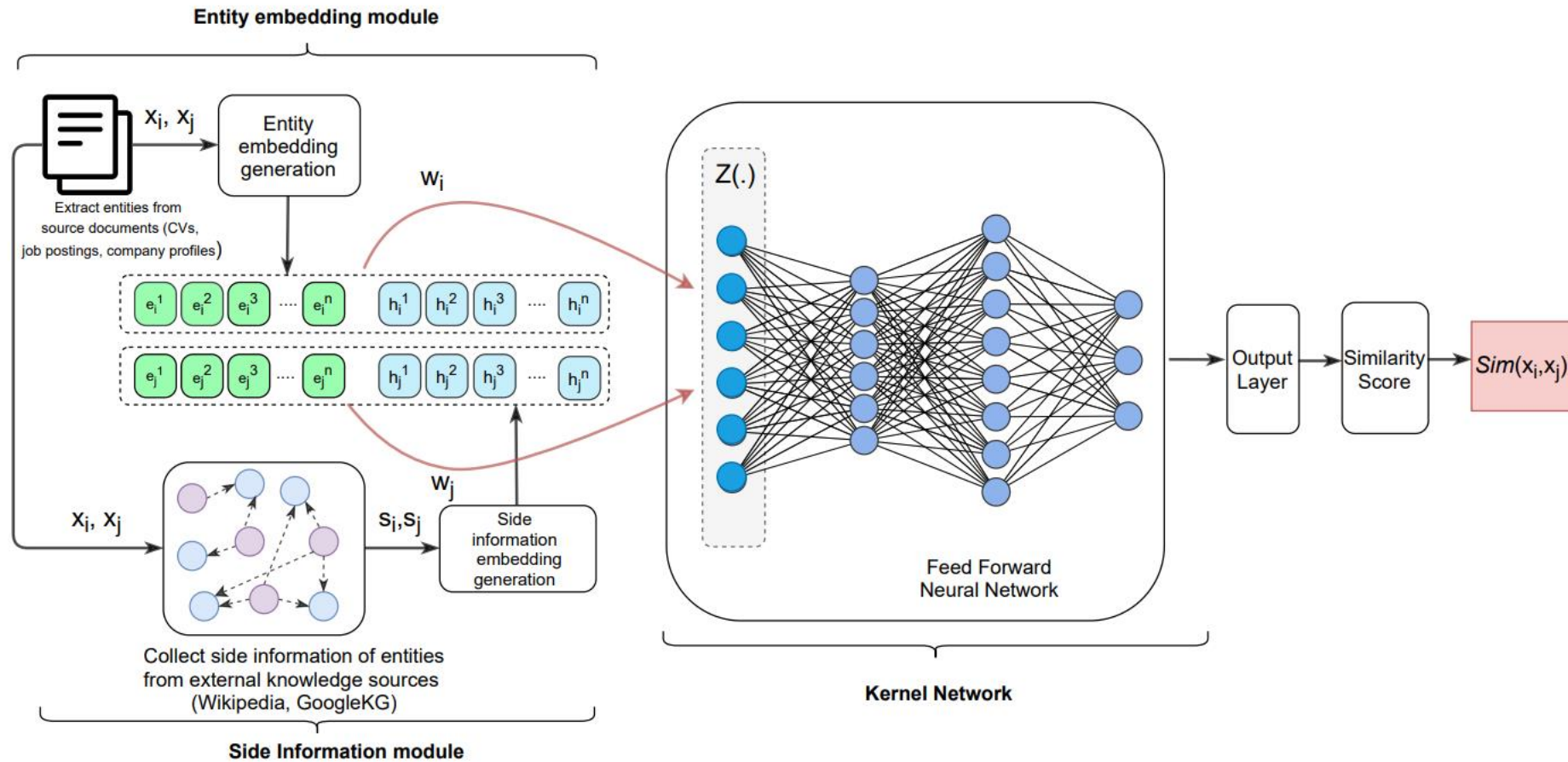
Side Information Collection from Wikipedia InfoBox and Google KG

Proposed Architecture



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

Proposed Framework

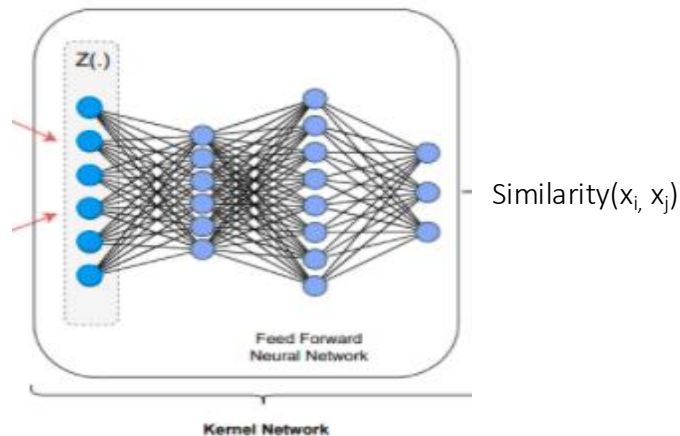


Proposed Framework

Z models element-wise relationships between input pairs.

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

$$Z = \{w_i^1 * w_j^1, \dots, w_i^{m+n} * w_j^{m+n}, |w_i^1 - w_j^1|, \dots, |w_i^{m+n} - w_j^{m+n}|\}$$



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

Results

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

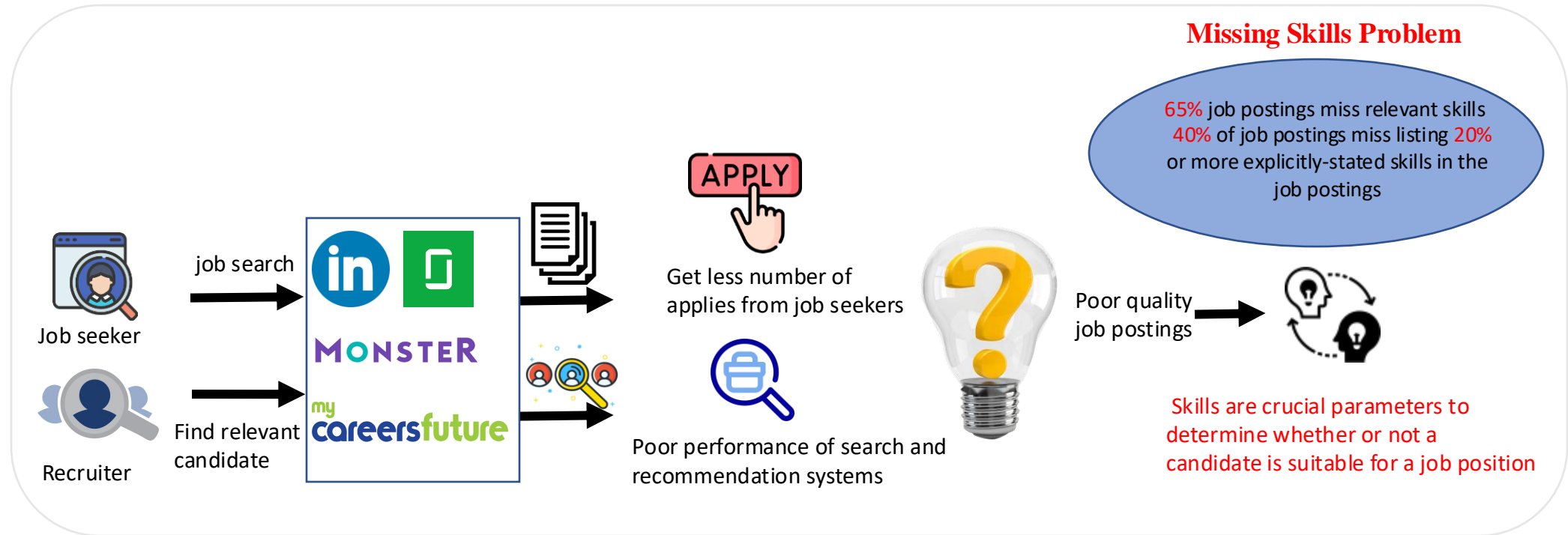
KCNet induces a non-linear mapping between the contextual vector representations while capturing fine-granular and high-dimensional relationships among vectors.

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.

Incomplete Content (Missing Skills)

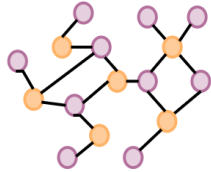


Finding Missing Skills

Job Title	Market Analyst					
Job description	Assist the Manager in sourcing the food industry and in conducting product research and analysis. Facilitate effective communication between the analytics and user experience teams. Evaluates customers' online behaviour and provide insights and recommendations for further enhancements to the guest experience. Strong research, data analysis and communication skills.					
Required skills	communication	data analysis	tableau	visualization	python	Excel
Explicit Skills			Implicit Skills			

An example of a job posted over a recruitment platform. The job description does not include implicit and job-specific skills such as 'tableau', 'visualization', 'python', and 'Excel'.

Finding Missing Skills



Novel job-skill graph consisting of 22, 844 nodes(jobs and skills) and 650K relationships



Formulated and proposed a framework, JobXMLC that learns a job-skill graph with multi-resolution **graph** neighborhoods



JobXMLC outperforms by a margin of 6% from the SOTA baselines



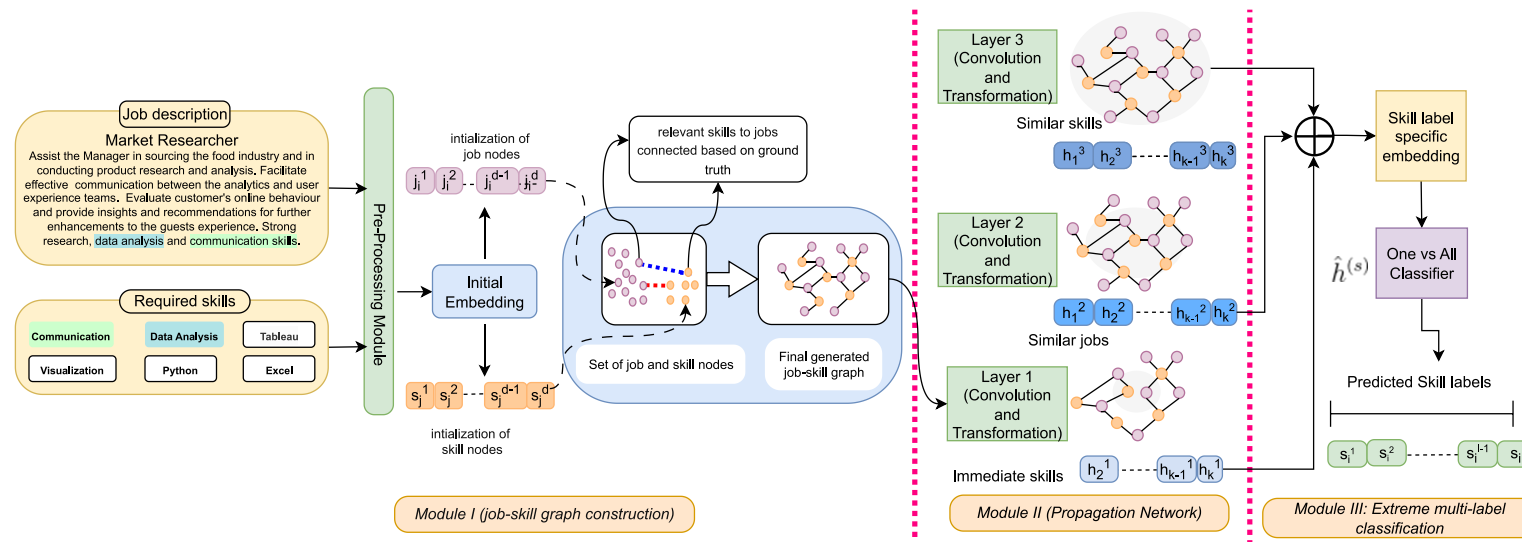
JobXMLC is lightweight, up to 18X faster in training and 634X in predicting than existing deep learning-based extreme classifiers

Dataset Statistics

Element	mycareersfuture.sg	StackOverflow Jobs
#No. of job posts	20, 298	20, 320
# of distinct skills	2, 548	275
# of skills with 20 or more mentions	1, 209	50
Average skill tags per job post	19.98	2.8
Average token count per job post	162.27	200.8
Maximum token count in a job post	1, 127	800

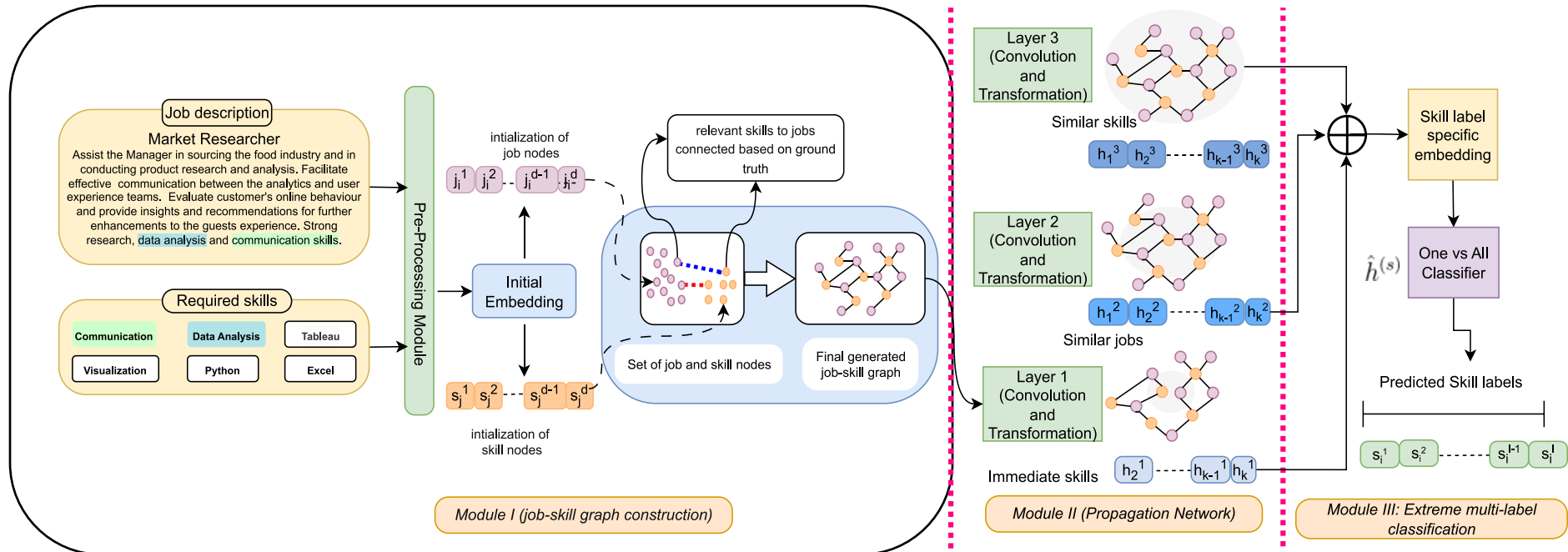
Dataset statistics for mycareersfuture.sg and StackOverflow Jobs

Proposed Approach



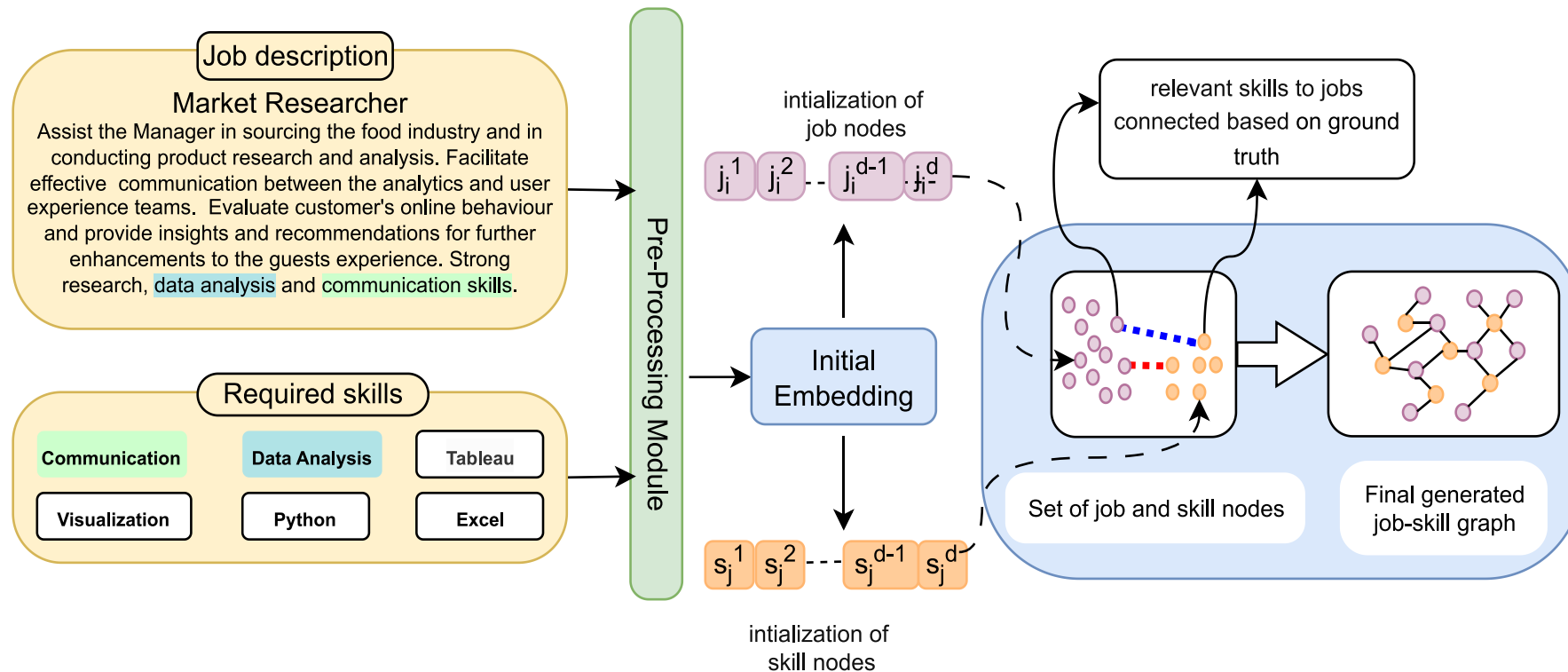
JobXMLC consists of three components: Module I consists of a mechanism to construct a job-skill graph, and Module II consists of a graph neural network-based architecture that learns embeddings using multi-hop neighborhoods using a job-skill graph effectively. Module III uses a scalable mechanism of extreme classifiers to predict missing skills.

Proposed Approach



JobXMLC consists of three components: Module I consists of a mechanism to construct a job-skill graph, and Module II consists of a graph neural network-based architecture that learns embeddings using multi-hop neighborhoods using a job-skill graph effectively. Module III uses a scalable mechanism of extreme classifiers to predict missing skills.

Job-skill graph construction

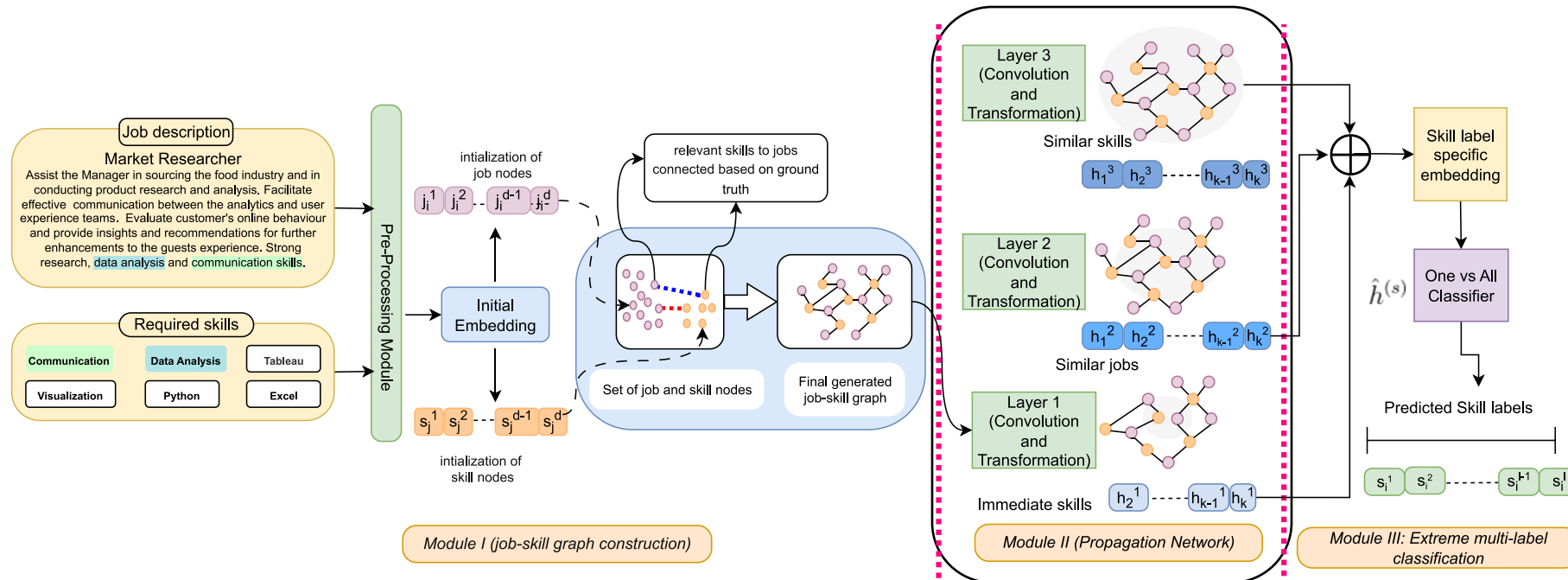


Pre-process job description (apply POS tagging to filter uninformative words)

Initialize the nodes (jobs, skills) using representations from a light-weight embedding model

Skills relevant to jobs are connected based on ground truth (required skills)

Proposed Approach



JobXMLC consists of three components: Module I consists of a mechanism to construct a job-skill graph, and Module II consists of a graph neural network-based architecture that learns embeddings using multi-hop neighborhoods using a job-skill graph effectively. Module III uses a scalable mechanism of extreme classifiers to predict missing skills.

Propagation Network

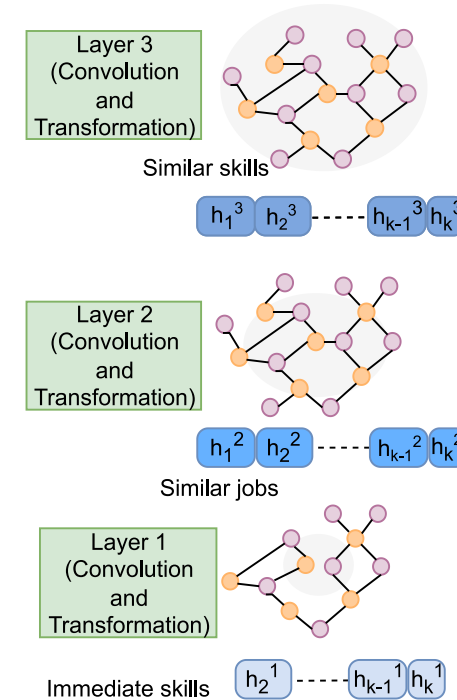
- Exploits higher-order job-skill graph structure using multiple layers of aggregation
- Convolution aggregates information from node neighbors
- Transformation update the node representation based on convolved embeddings
- The network is able to capture transitive cues if jobs j_1 and j_2 share a common skill s_1 . Another skill s_2 relevant to j_2 , we infer that s_2 might also be relevant to j_1

$$f_v^{(k)} = (1 + \lambda_k) f_v^{(k-1)} + \sum_{j \in \mathcal{N}_v, j \neq v} f_j^{(k-1)} \quad (1)$$

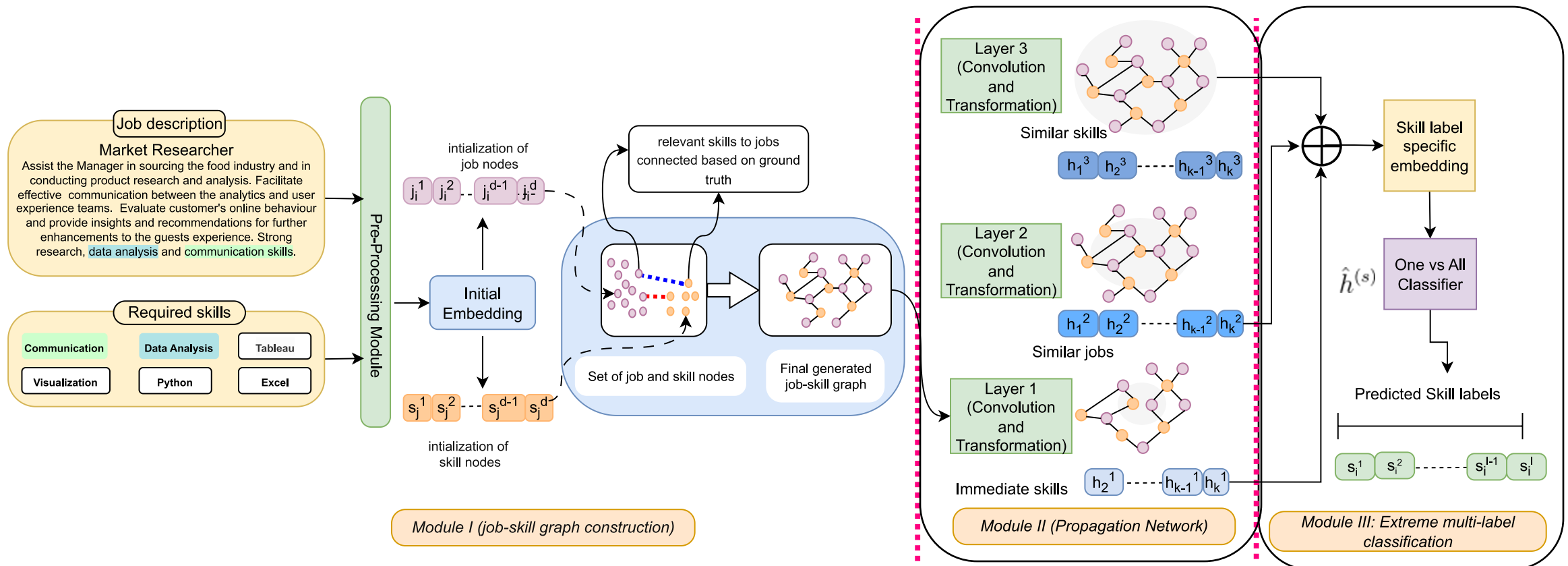
where \mathcal{N}_v be the set of neighboring nodes of an i^{th} node; $f_v^{(k)}$ be the representation of the v^{th} node after layer k , and λ is a fixed scalar for layer k .

$$h_v^{(k)} = f_v^{(k)} + g(\delta(R_k * g(f_v^{(k)}))) \quad (2)$$

where $g(\cdot)$ is ReLU activation, $\delta(\cdot)$ is batch normalization and R_k is a parameter matrix for the residual layer.



Proposed Approach

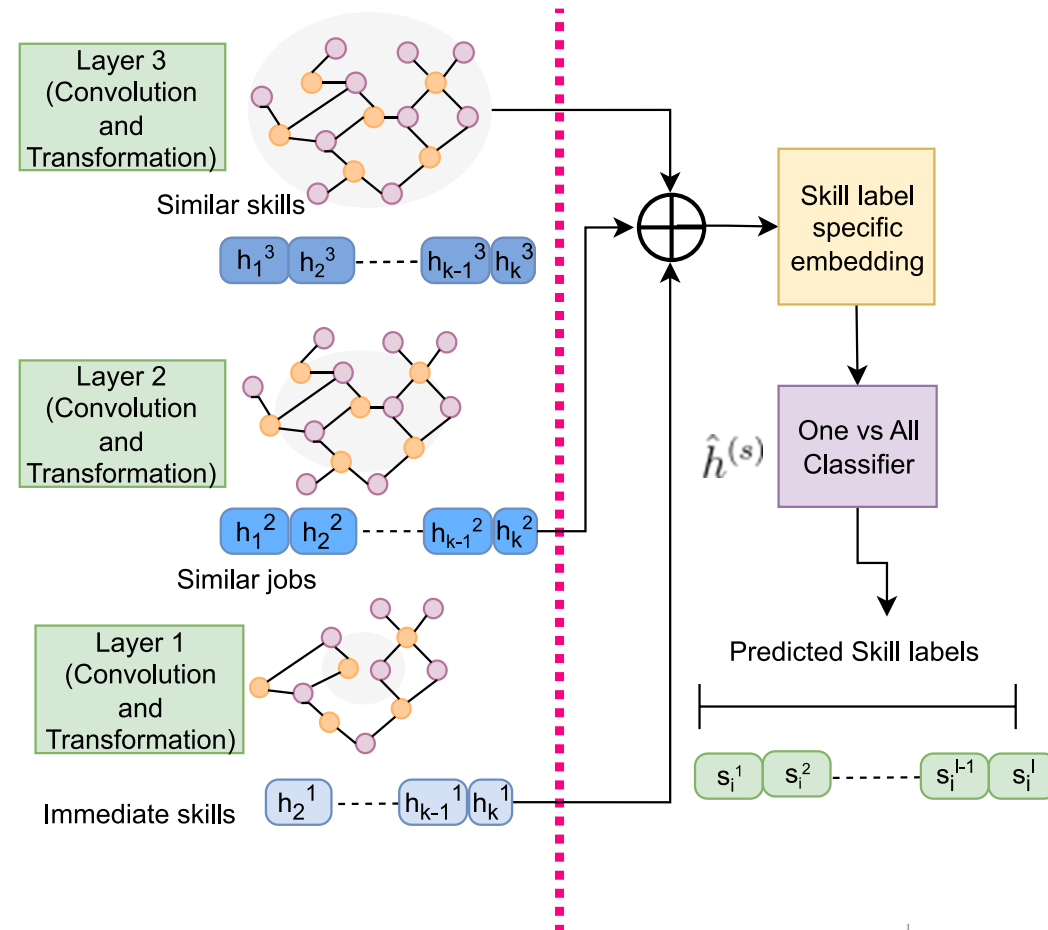


JobXMLC consists of three components: Module I consists of a mechanism to construct a job-skill graph, and Module II consists of a graph neural network-based architecture that learns embeddings using multi-hop neighborhoods using a job-skill graph effectively. Module III uses a scalable mechanism of extreme classifiers to predict missing skills.

Extreme multi-label classification

Incorporate label-wise attention for every skill and obtain their attention weights

Calculate score for label-specific embedding and then
One vs all classifier is used to obtain a score for a skill label



Qualitative Analysis

Job description	minimum 5 7 years experience information technology software development must 3 4 yeras experience dot net development experience asp.net c .net xml experience. language query update etc knowledge pc networking require dot net developer mnc client singapore typre position long term contract initial degree information technology require minimum 5 7 years experience information technology software development must 3 4 years experience dot net development experience asp.net c net xml etcknowledge pc networking good communication skills									
Required skills (Ground truth)	Software development	java	.NET	Javascript	jQuery	XML	Web applications	ASP.NET	SDLC	
BERT-XMLC+CAB	Software development	java	.NET	jQuery	XML	PHP	Python	C	Linux	Software engineering
JOBXMLC	Software development	java	.NET	Javascript	jQuery	XML	Web applications	ASP.NET	SDLC	integration

Shows the skills predicted by BERT–XMLC+CAB and JobXMLC where input is job description. **Purple** shows correct skill predictions by JobXMLC as compared with required skills (ground truth). **Green** shows the extra skills predicted by JobXMLC. **Red** skills are missed by BERT+XMLC+CAB model as compared with ground truth.

Experiments and Results

Model	R@5	R@10	R@30	P@5	P@10	P@30
CNN	14.17	23.58	45.34	56.67	47.17	30.23
LSTM [†]	11.67	18.44	35.02	46.67	36.89	23.34
Bi-LSTM [†]	13.02	21.37	41.54	52.07	42.75	27.70
Bi-GRU [†]	13.98	23.43	44.41	55.94	46.87	29.61
BERT+XMLC	15.27	25.96	51.18	61.06	51.92	39.32
RoBERTa+XMLC	16.15	26.52	51.99	60.08	53.85	39.87
BERT+XMLC+CAB	16.72	29.45	58.98	66.87	58.90	41.21
GalaXC	16.31	28.34	54.16	65.25	56.70	36.11
JobXMLC (GraphSaint)	16.23	27.79	53.32	64.93	55.59	35.55
JobXMLC (GraphSAGE)	16.84	29.18	56.89	67.36	58.36	37.93
JobXMLC	18.29	32.33	63.18	73.20	64.66	42.22

Results of JobXMLC along with state-of-the-art approaches on mycareersfuture.sg dataset. For RNN-based models (†), we have limited all model architectures to two layers.

Experiments and Results

Model	R@5	R@10	R@30	P@5	P@10	P@30
CNN	25.16	39.39	64.80	15.24	11.72	6.36
LSTM [†]	26.63	40.47	67.89	16.07	11.95	6.65
Bi-LSTM [†]	41.46	55.27	76.38	23.83	16.12	7.56
Bi-GRU [†]	46.15	59.01	78.61	26.68	17.23	7.79
BERT+XMLC	35.50	50.95	76.06	20.75	14.99	7.58
RoBERTa+XMLC	36.20	52.23	77.05	21.98	15.09	7.88
BERT+XMLC+CAB	37.20	51.24	78.98	22.18	15.02	8.03
GalaXC	43.27	51.47	67.50	24.23	14.53	6.50
JobXMLC (GraphSaint)	39.16	51.73	73.99	22.28	14.88	7.22
JobXMLC (GraphSAGE)	38.76	52.26	74.19	21.98	14.99	7.23
JobXMLC	47.85	59.26	74.53	26.92	16.94	7.23

Table 3: Results of JobXMLC along with state-of-the-art approaches on StackOverflow Jobs dataset. For RNN-based models ([†]), we have limited all model architectures to two layers.

Performance Comparison

Datasets →	mycareersfuture.sg		StackOverflow Jobs	
Models ↓	Training Time (in hours)	Prediction Time (in ms)	Training Time (in hours)	Prediction Time (in ms)
BERT+XMLC	5.50	1200	1.63	350
RoBERTa+XMLC	4.72	1200	1.24	350
BERT+XMLC+CAB	9.50	1200	4.86	350
JobXMLC	0.51	1.89	0.31	1.71

Comparison of JobXMLC with stronger baselines. JobXMLC is faster to train than leading Deep Extreme Classifiers like BERT at training time and prediction time.

Summary

Proposed a JobXMLC framework, which uses a graph neural network to incorporate neighborhood information with the help of a collaborative graph over jobs and skills

JobXMLC leverages skill attention mechanism and attends to multi-resolution representations of jobs and skills

JobXMLC outperforms leading deep extreme classifiers on precision and recall metrics by 6% and 3% respectively

JobXMLC is 18X faster on training and 634X faster on predicting than deep extreme classifiers and can be scaled efficiently to real-world datasets with thousands of labels

3. Con2KG FRJD

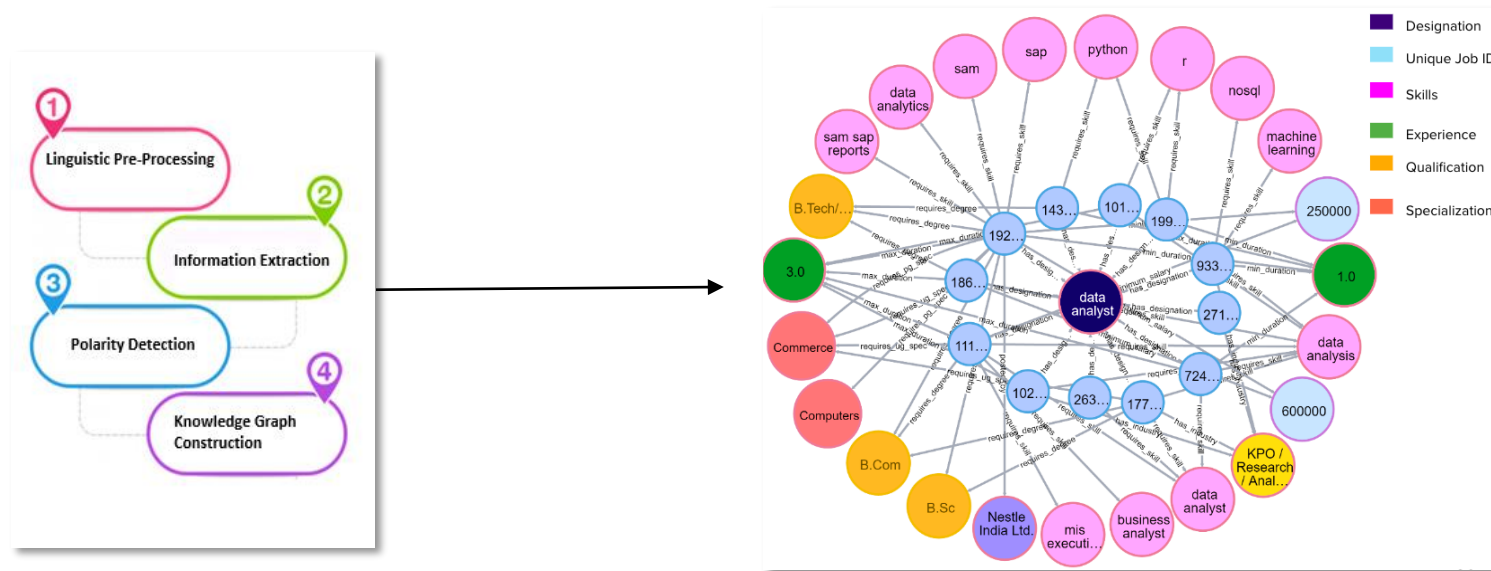
Built a Domain-specific Knowledge Graph (Con2KG) and developed a novel framework FRJD to classify fraudulent and legitimate jobs

Identify misleading content

Extract domain-specific information from job postings and construct domain-specific knowledge base ([Building the Domain-Specific Knowledge Graphs](#))

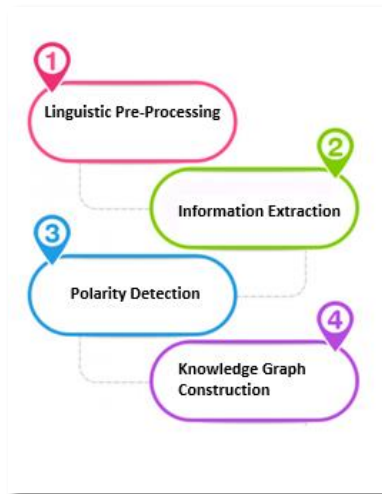
Build a framework to classify misleading content using domain knowledge

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.*

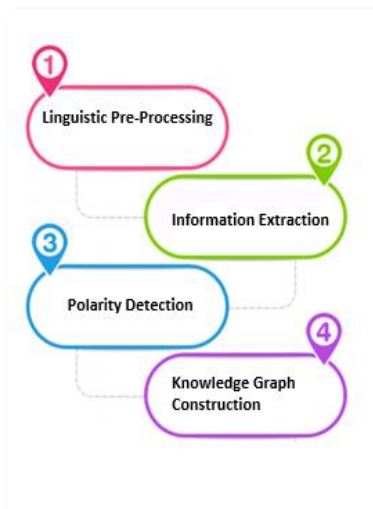
Building the Domain-Specific Knowledge Graphs



- A. Preprocess the noisy, unstructured and semi-structured data from job postings using NLP techniques
- B. To accomplish this task, we
 - A. Employed **sentence detection module**
 - B. Revived **missing phrases** using POS Tagging
 - C. Removed **HTML Non-ASCII** characters.
- C. 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.*

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.



Companies



Institutes



Skills



Experience



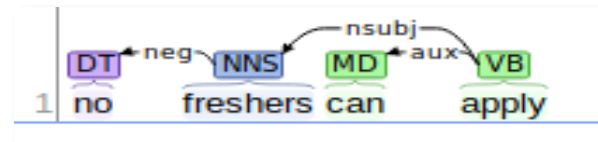
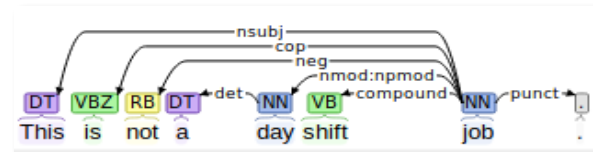
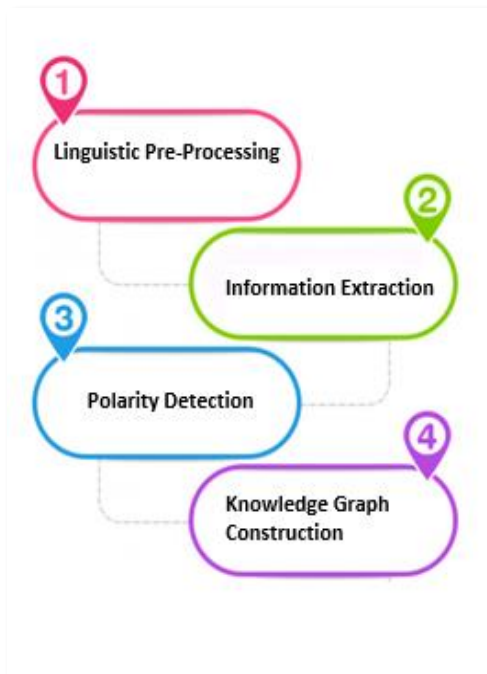
Certifications



Designations

*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.*

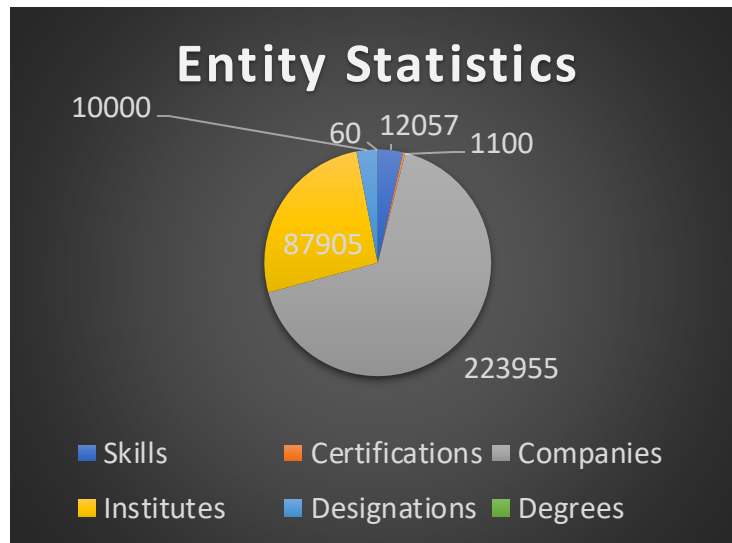
Building the Domain-Specific Knowledge Graphs



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.

Building the Domain-Specific Knowledge Graphs



250,000 Job postings

5,220 unique relations

linking 3,65,0,61 Entities

40,11,030 relationships

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

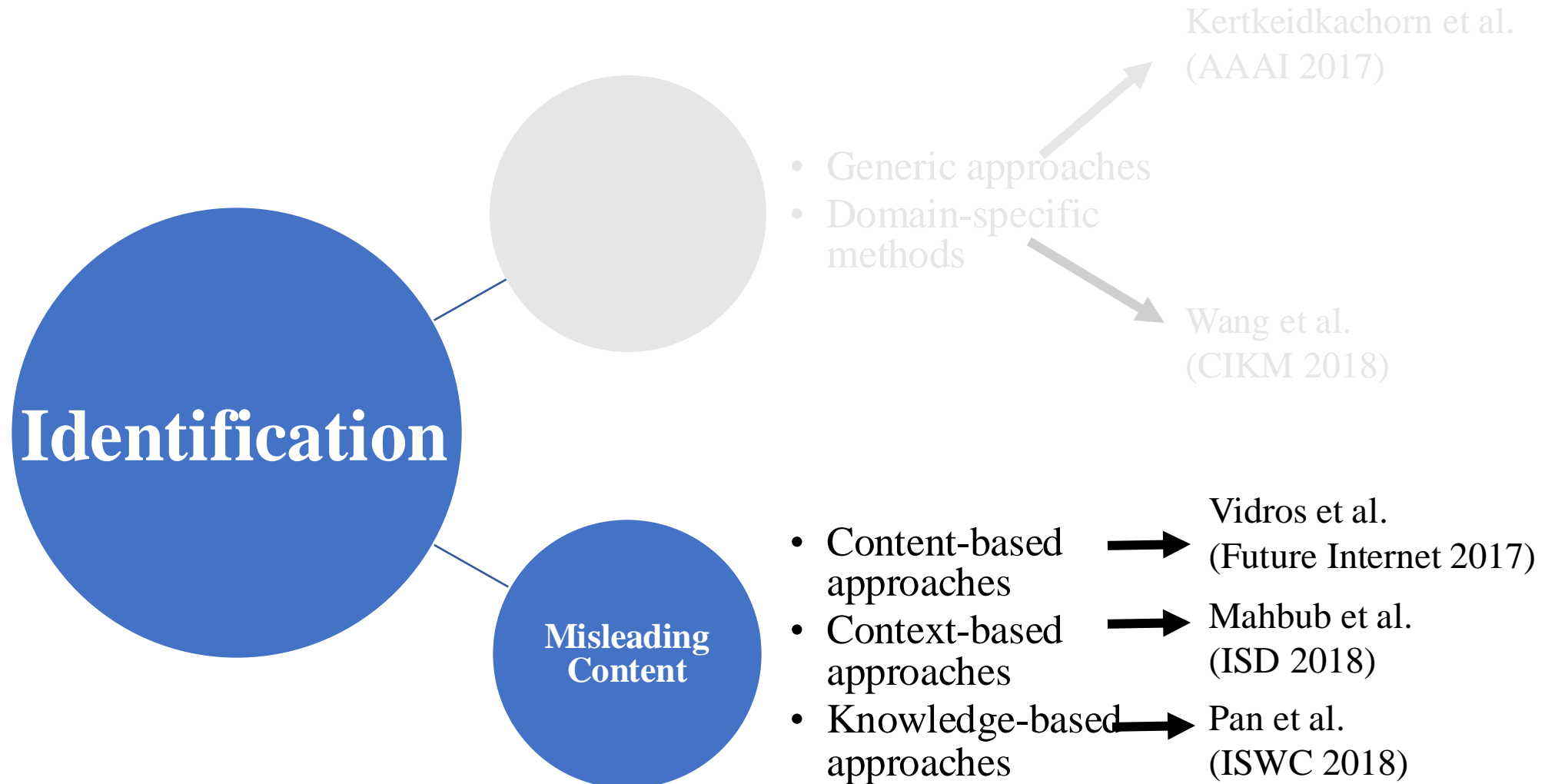
Misleading content

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

<p>Data Entry Clerks Position</p> <p>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.</p>	<p>Data Entry Clerk</p> <p>Responsibilities include, but are not limited to:</p> <ul style="list-style-type: none"> 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



Research Gap

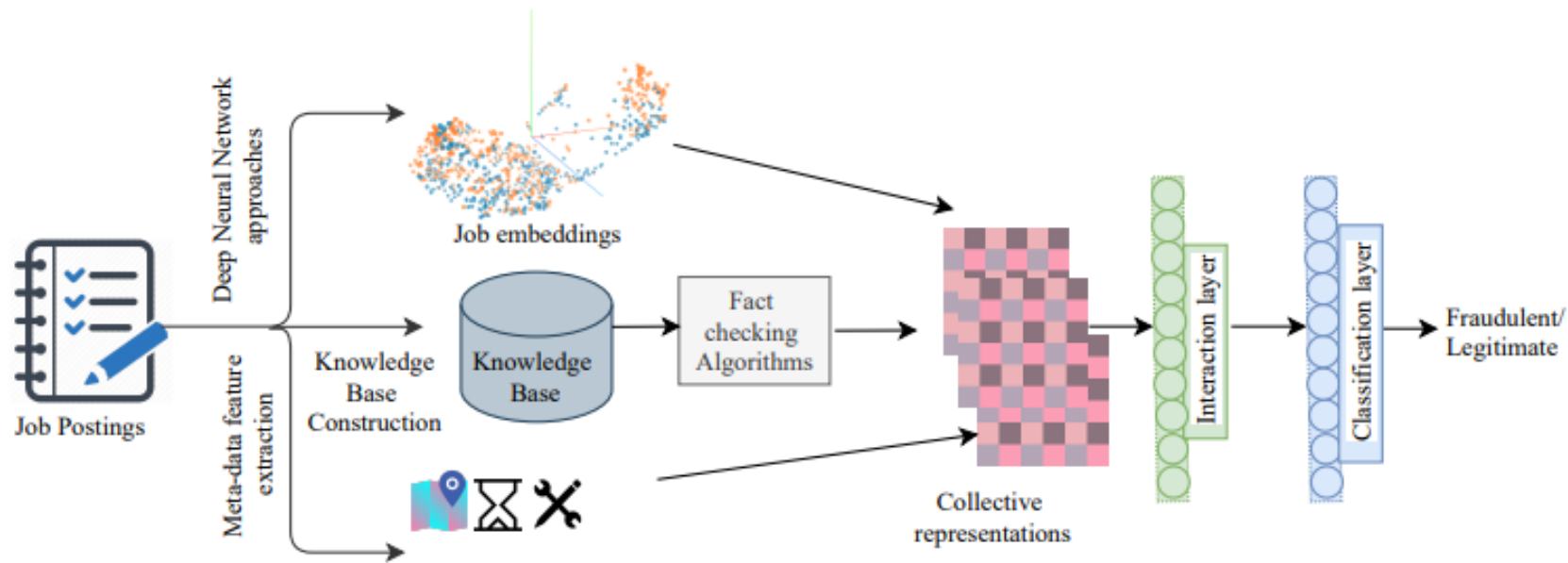
Handcrafted , linguistic, writing styles, string-based features.
Ignore the factual information among domain-specific entities
present in job postings.

Identification

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)

Proposed Framework



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).

Objective

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

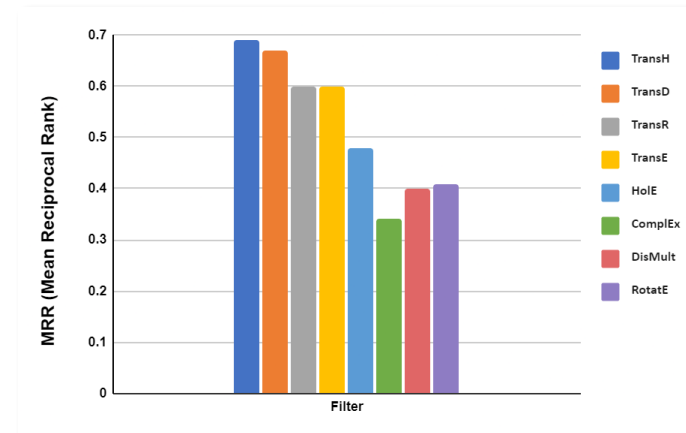
$KGA \in \{TransE, TransR, TransH, TransD, DistMult, ComplEx, HolE, RotatE\}$

Results

MRR (Mean Reciprocal Rank) metric for triple prediction

TransH outperforms the other fact-checking algorithms

TransH is able to model many-to-many relationships well



Results

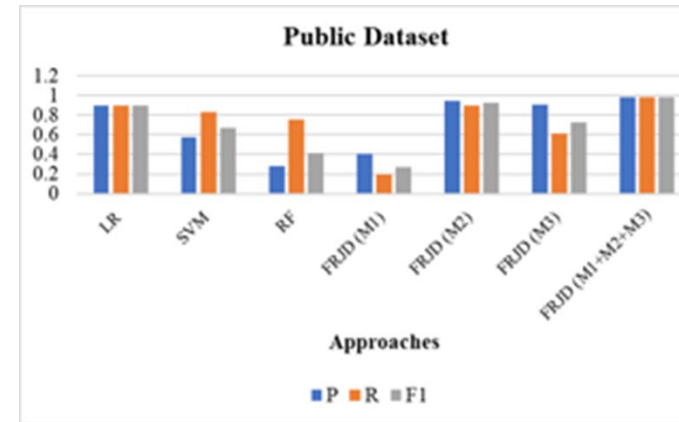
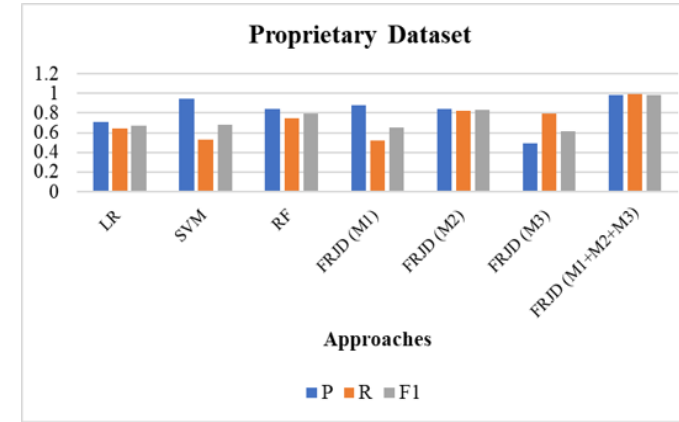
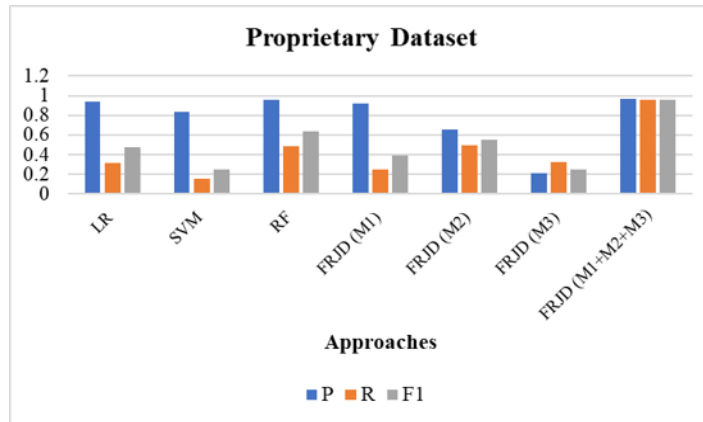


Fig. 3. Evaluation results on proprietary and public datasets for job postings a) fraudulent class and b) legitimate class at the right where M1, M2, M3 are contextual, factual, and meta features.

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).



JobXMLC: EXtreme Multi-Label Classification of Job Skills with Graph Neural Networks

Nidhi Goyal
IIIT-Delhi
Raghava Mutharaju
IIIT-Delhi

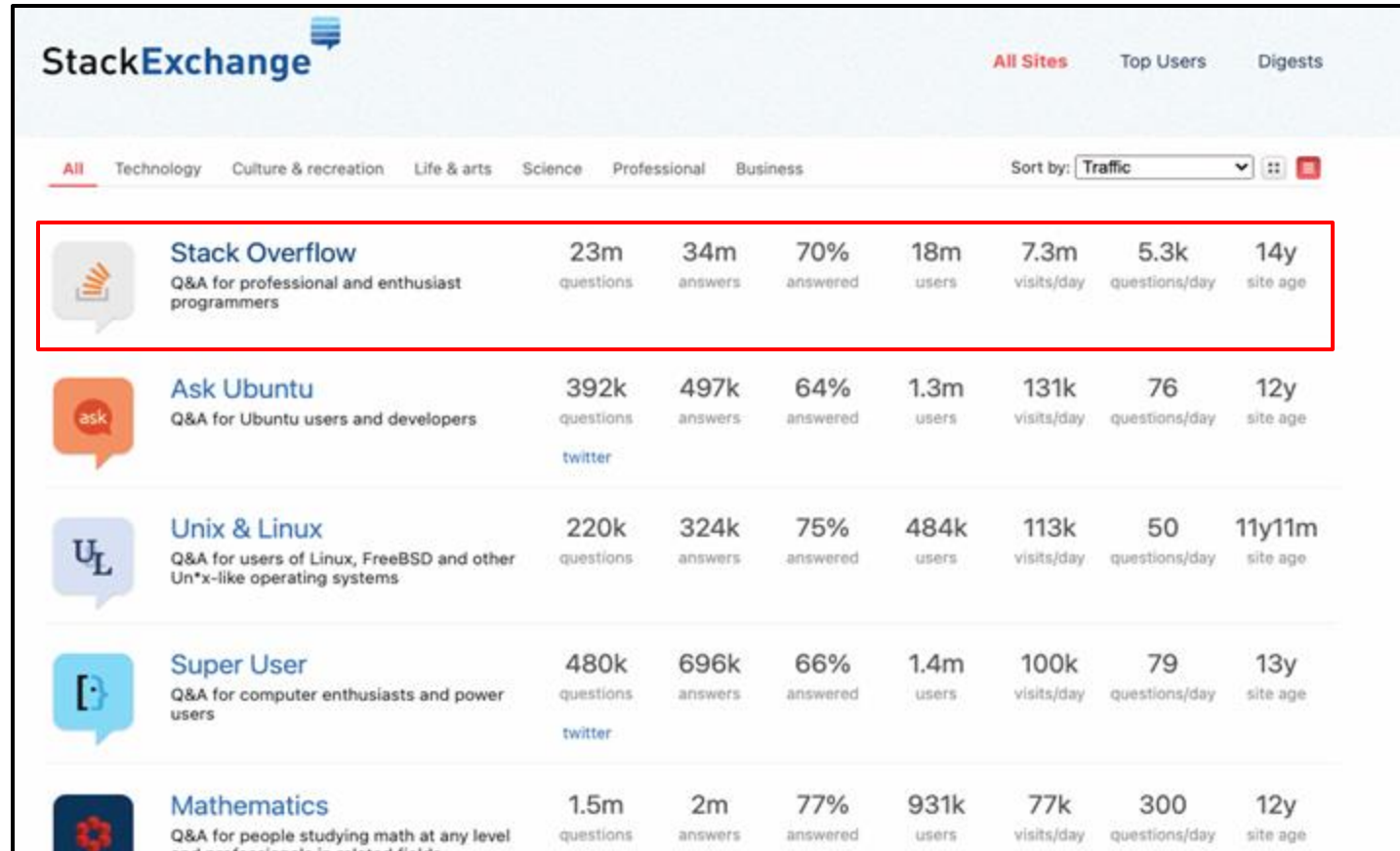
Jushaan Singh Kalra
DTU, Delhi
Niharika Sachdeva
InfoEdge India Limited

Charu Sharma
IIIT-Hyderabad
Ponnurangam Kumaraguru
IIIT-Hyderabad

4. LQuaD

Developed an architecture to identify low quality (off-topic, too broad, opinion-based, unclear what are you asking) questions and






Identify low-quality content



StackExchange

All Sites Top Users Digests

All Technology Culture & recreation Life & arts Science Professional Business Sort by: Traffic

	Stack Overflow Q&A for professional and enthusiast programmers	23m questions	34m answers	70% answered	18m users	7.3m visits/day	5.3k questions/day	14y site age
	Ask Ubuntu Q&A for Ubuntu users and developers	392k questions	497k answers	64% answered	1.3m users	131k visits/day	76 questions/day	12y site age
	Unix & Linux Q&A for users of Linux, FreeBSD and other Un*x-like operating systems	220k questions	324k answers	75% answered	484k users	113k visits/day	50 questions/day	11y11m site age
	Super User Q&A for computer enthusiasts and power users	480k questions	696k answers	66% answered	1.4m users	100k visits/day	79 questions/day	13y site age
	Mathematics Q&A for people studying math at any level	1.5m questions	2m answers	77% answered	931k users	77k visits/day	300 questions/day	12y site age

Identify low-quality content

Stack Overflow > Help center > Asking

How do I ask a good question?

We'd love to help

[Search](#), and

...and keep track
links to related c
different from th

Write a title

The title is the first thing potential answerers will see, and if your title isn't interesting, they won't read the rest. *So make it count:*

- **Pretend you're talking to a busy colleague** and have to sum up your entire question in one sentence: what details can you include that will help someone identify and solve your problem? Include any error messages, key APIs, or unusual circumstances that make your question different from similar questions already on the site.
- **Spelling, grammar and punctuation are important!** Remember, this is the first part of your question others will see - you want to make a good impression. If you're not comfortable writing in English, ask a friend to proof-read it for you.
- If you're having trouble summarizing the problem, **write the title last** - sometimes writing the rest of the question first can make it easier to describe the problem.

Include all relevant tags

Try to include a tag for the language, library, and specific API your question relates to. If you start typing in the tags field, the system will suggest tags that match what you've typed - be sure and read the descriptions given for them to make sure they're relevant to the question you're asking! See also: [What are tags, and how should I use them?](#)

Low-quality content



Title	Which strategy project?	Title	Convert a list to dict key value
Body	You work on an important project that contains 7 independent choose?	Body	What I have tried so far is: self.dict_total_words1 = {i.split(': ')[0]: int(i.split(': ')[1]) for i
Tags	time-management	Tags	python regex java
Decision	Closed: off-topic	Decision	Closed: unclear what you're asking
Title	Why do people hate java?	Title	How do I code a forum in PHP?
Body	In the world of python programming	Body	Hi, I'm a beginner. I've been tasked with coding a web forum in PHP
Tags	java python	Tags	php sql
Decision	Closed: primarily opinion-based	Decision	Closed: too broad

Fig. 1. A sample of 'closed' questions from Stack Overflow. These are 'closed' due to different reasons such as 'off-topic', 'unclear what you're asking', 'too broad' and 'primarily opinion-based'.

Contributions

Propose a novel framework, **LQuaD**, which establishes the utility of a question-tag graph and transformers to detect low-quality questions that are likely to get ‘closed’ at the time of posting. Our framework acts as an early-assessment tool to assist users in composing a question, which would remain open and receive responses.

Examine the impact of non-content related characteristics of the question using survival analysis to estimate the time duration of closure of the question.

Evaluate **LQuaD** on dataset of ‘closed’ and non-‘closed’ questions from Stack Overflow platform and make the code publicly available for reproducibility.

Contributions

Propose a novel framework, LQuaD, which establishes the utility of a question-tag graph and transformers to detect low-quality questions that are likely to get 'closed' at the time of posting. Our framework acts as an early-assessment tool to assist users in composing a question, which would remain open and receive responses.

Examine the impact of non-content related characteristics of the question using survival analysis to estimate the time duration of closure of the question.

Evaluate LQuaD on dataset of 'closed' and non-'closed' questions from Stack Overflow platform and make the code publicly available for reproducibility.

Problem Formulation

Consider the set of questions $Q = \{q_1, q_2, \dots, q_i\}$, a question $q_i \in Q$ is a tuple (c_i, T_i, y_i) where c_i, T_i, y_i represents the content, tag set, and label of the i^{th} question.

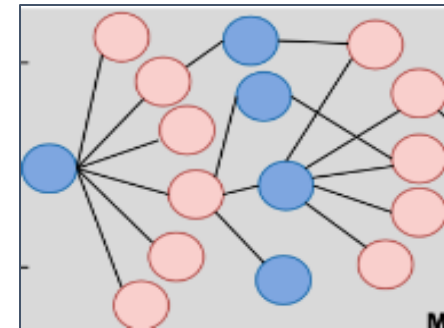
Content (c_i) : *Title* \oplus *Body*

Tag set (T_i) : $\{t_{i1}, t_{i2}, t_{i3} \dots t_{ik}\} \forall 1 \leq k \leq 6$ and t_{ik} represents the k^{th} tag of the i^{th} question.

Class (y_i) : $y_i = 0 \rightarrow \text{closed}$ $y_i = 1 \rightarrow \text{non-closed}$

We construct an undirected question – tag graph $G = (V, E)$

$V \Rightarrow Q$ and $T = \{T_1 \cup T_2 \cup \dots \cup T_{|Q|}\}$
 $E \Rightarrow q_i - t_{ik}$ and $E \subset Q \times T$



Proposed Approach

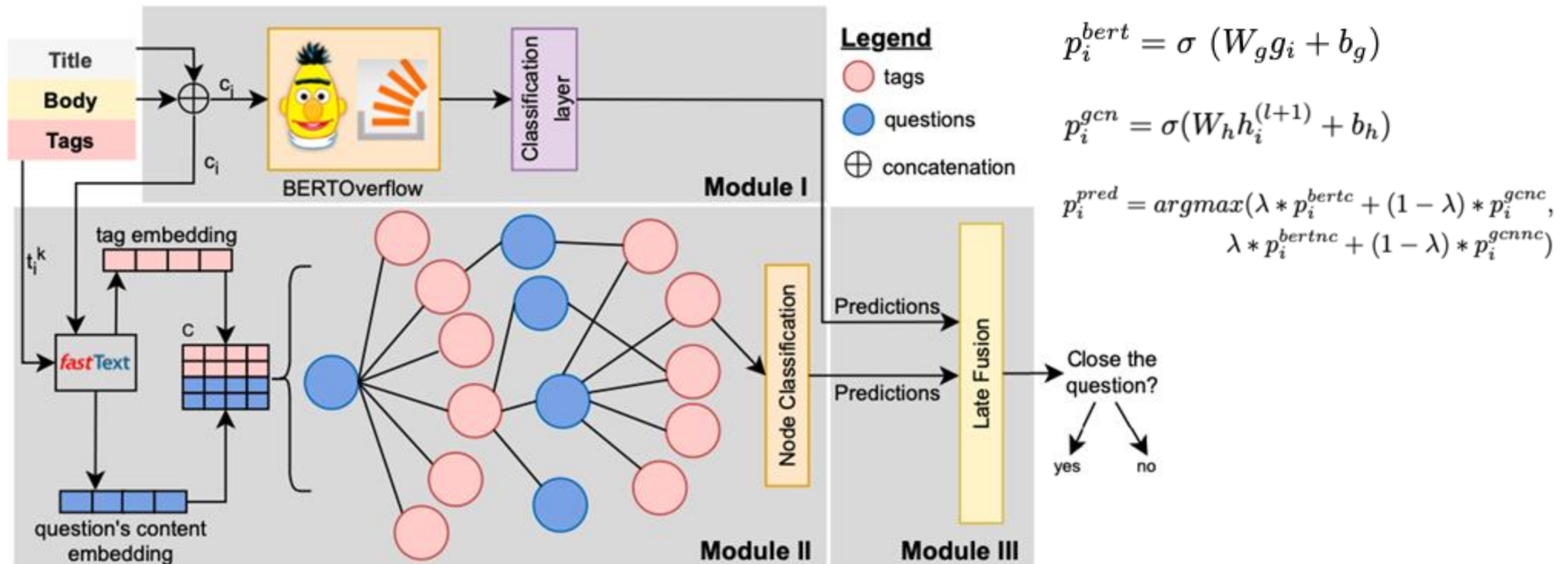


Fig. 2. LQuAD consists of three components: Module I fine-tunes the BERTOverflow model, which inputs the title and body for the question's classification task, Module II consists of a graph convolutional network initialized with the question's content and tag embedding for the node classification task, Module III consists of a late fusion strategy that combines predictions from both modules.

Proposed Approach

Propose a novel framework, **LQuaD**, which establishes the utility of a question-tag graph and transformers to detect low-quality questions that are likely to get 'closed' at the time of posting. Our framework acts as an early-assessment tool to assist users in composing a question, which would remain open and receive responses.

Examine the impact of non-content related characteristics of the question using survival analysis to estimate the time duration of closure of the question.

Evaluate **LQuaD** on dataset of 'closed' and non-'closed' questions from Stack Overflow platform and make the code publicly available for reproducibility.

Temporal Event Analysis

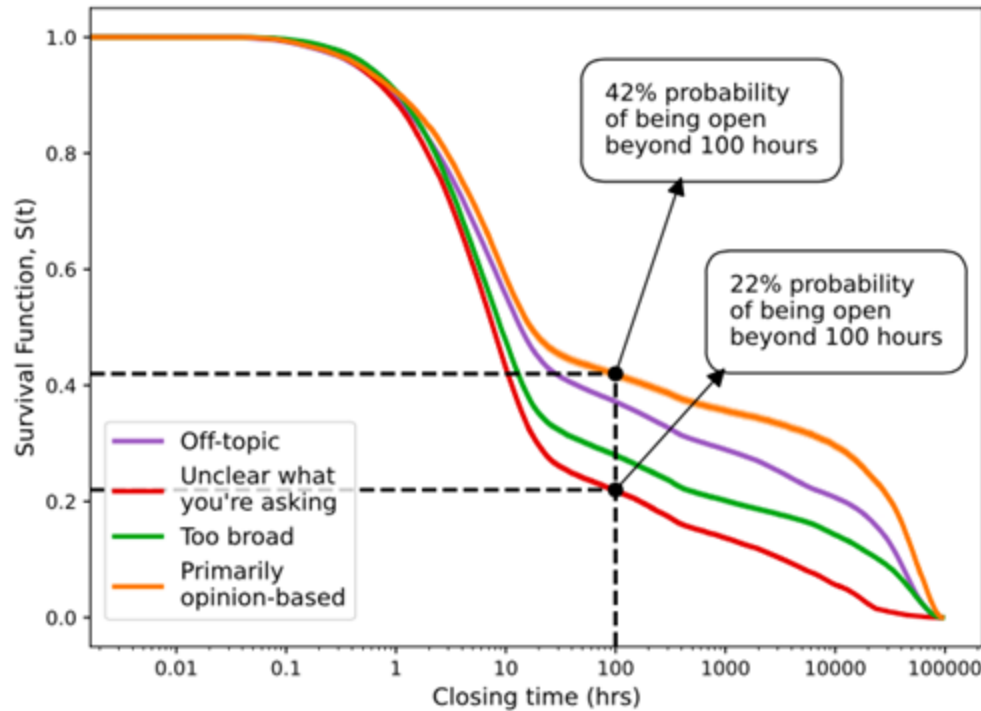
	Reasons				Tag Categories				
Category	Off-topic	Unclear what you're asking	Too broad	Primarily opinion-based	Database	Cloud	Web frameworks	Programming languages	Other frameworks
Mean Time Till EOI (days)	330.76	<u>65.91</u>	230.87	531.81	<u>145.54</u>	228.13	185.03	177.89	482.89
Median Survival Time (hrs)	12.89	<u>7.51</u>	8.57	15.07	9.04	27.33	8.17	<u>7.27</u>	16.51

Table I : We report the mean time (days) till EOI and median survival time (hrs) corresponding to reasons and tag categories. The highest values in respective rows are shown in bold and the lowest values are underlined.

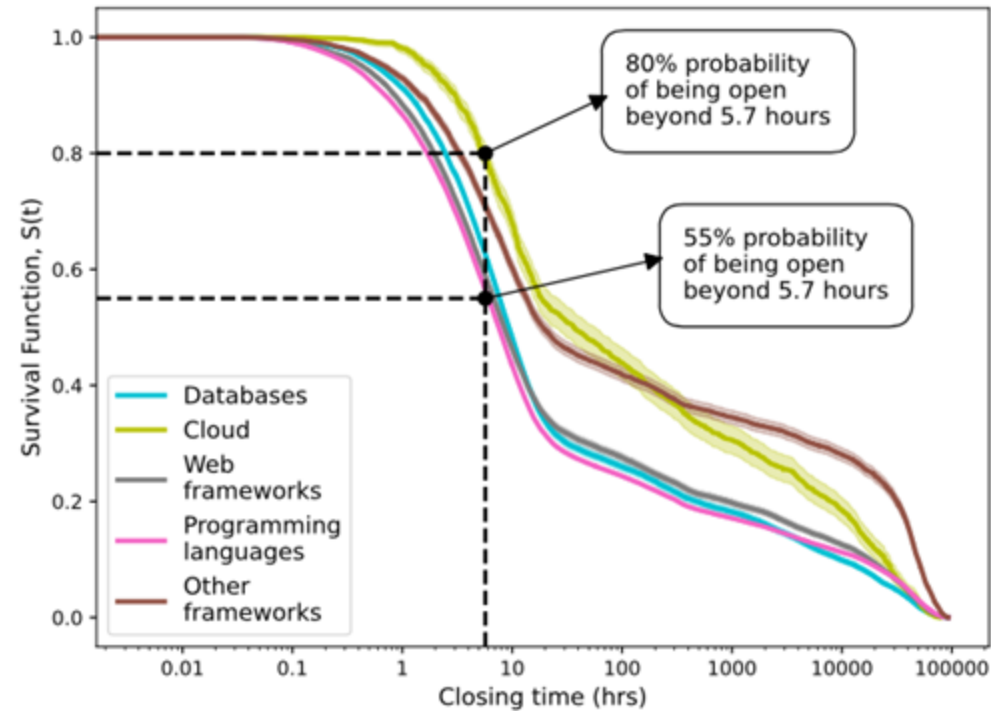
'primarily opinion-based questions' remain open for longer time duration than *'unclear what you're asking'* questions.

Question with tags in *Cloud* and *other frameworks* have higher mean time till EOI and median $\hat{S}(t)$ than other categories and are not closed for relatively longer time periods.

Temporal Event Analysis



(a)



(b)

Fig. 3. Kaplan-Meier estimator of survival function for time period for 'closed' questions based on different reasons of 'closing'. Plot (a) specifies the tag category whereas Plot (b) specifies the reasons due to which the question get 'closed'.

Dataset Statistics

	Module I			Module II	
Dataset	No. of train samples	No. of val. (test) samples	Total	Graph	Total
'Closed'	155,554	51,852	259,258	Nodes (questions)	2,851,838
Non-'Closed'	1,555,548	518,516	2,592,580	Nodes (unique tags)	48,374
Total questions	1,711,102	570,368	2,851,838	Edges	8,442,584

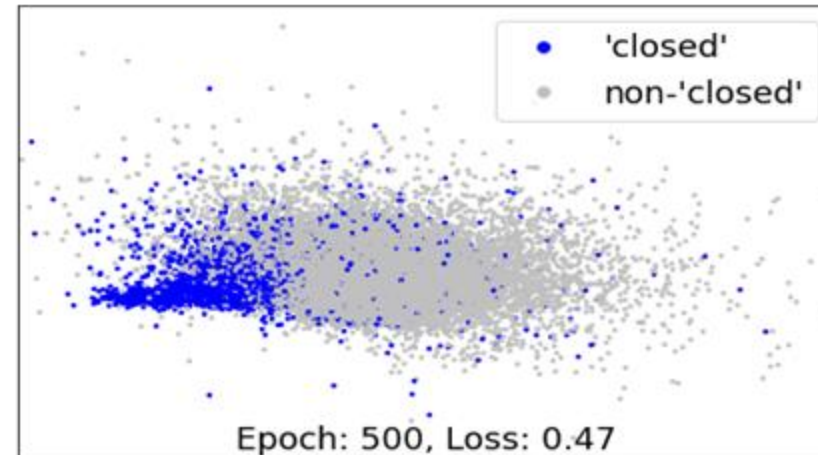
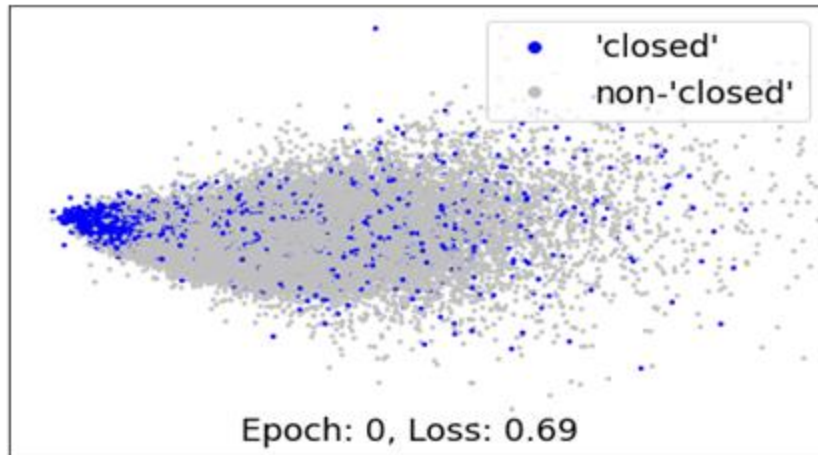
Table II : Dataset Statistics from the Stack Overflow Platform

Experimental Results

Model	Precision	Recall	F1
Denzil et al.	70.25	70.25	70.24
Toth et al.	73.78	73.33	73.66
Count Vectorizer + LR	89.94	76.90	81.38
Count Vectorizer + XGBoost	90.17	77.48	81.82
FastText + LR	90.52	79.71	83.44
DistilBERT +LR	92.19	85.26	87.59
GCN (fastText)	90.69	83.98	86.42
LQuaD (LF [mean])	94.85	95.20	94.86

Table III: Average performance metrics (weighted) on the Stack Overflow Dataset

Visualizations



Ablation Studies

Model	Precision	Recall	F1
Module I	94.81 (0.03)	95.08 (0.04)	94.53 (0.01)
Module II	91.86 (0.07)	84.38 (0.07)	86.92 (0.08)
LQuaD (LF [max])	93.62 (0.02)	92.85 (0.04)	93.16 (0.07)
LQuaD (LF [mean])	94.85 (0.03)	95.20 (0.05)	94.86 (0.02)

Table 4: Effectiveness of **LQuaD** (and variances) as compared to Module I and Module II

Conclusion

We propose **LQuaD** that incorporates semantic information of questions associated with each post using transformers and learns question and tag graphs in a transductive manner using GCNs. Our graph consists of 2.9M nodes and 8.4M edges. **LQuaD** detects low-quality questions that are likely to get 'closed' at the time of posting.

Our framework acts as an early-assessment tool to assist users in composing a question, which would remain open and receive responses.

We use survival analysis that reduces the number of questions close by informing users to take appropriate action.

LQuaD outperforms the state-of-the-art methods by a 21% in F1-score on the dataset of 2.8 million questions.

Summary

Developed a multi-tier framework, KCNET to normalize domain-specific entities (skills, institutes, companies, and designations)

Construction of a novel job-skill graph consisting of 22,844 (jobs and skills) and 650K relationships and novel framework, JobXMLC to find missing entities to improve the quality of jobs

Construction of Knowledge Graph (Con2KG) and Developed an architecture FRJD to classify fraudulent and legitimate jobs on online professional platforms

Proposed LQuaD that incorporates semantic information of questions associated with each post using transformers and learns question and tag graphs in a transductive manner using GCNs.

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.

Publications

1. **Goyal, N.**, Kalra, J., Sharma, C., Mutharaju, R., Sachdeva, N., & Kumaraguru, P. (2023, May). JobXMLC: EXtreme Multi-Label Classification of Job Skills with Graph Neural Networks. In Findings of the Association for Computational Linguistics: EACL 2023 (pp. 2136-2146).
2. **Goyal, N.**, Mamidi, R., Sachdeva, N., & Kumaraguru, P. (2023, January). Warning: It's a scam!! Towards understanding the Employment Scams using Knowledge Graphs. In Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD) (pp. 303-304).
3. **Goyal, N.**, Arora, U., Goel, A., Sachdeva, N., & Kumaraguru, P. (2022, July). Ask it right! Identifying low-quality questions on community question answering services. In 2022 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.
4. **Goyal, N.**, Sachdeva, N., Goel, A., Kalra, J. S., & Kumaraguru, P. (2021, September). KCNet: Kernel-based canonicalization network for entities in recruitment domain. In Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part II 30 (pp. 157-169). Springer International Publishing.

Publications

5. **Goyal, N.**, Sachdeva, N., & Kumaraguru, P. (2021, August). Spy the lie: fraudulent jobs detection in recruitment domain using knowledge graphs. In Proceedings of Knowledge Science, Engineering and Management: 14th International Conference, KSEM 2021, Part II 14 (pp. 612-623). Springer International Publishing.

6. **Goyal, N.**, Sachdeva, N., Choudhary, V., Kar, R., Kumaraguru, P., & Rajput, N. (2019, September). Con2kg-a large-scale domain-specific knowledge graph. In Proceedings of the 30th ACM Conference on Hypertext and Social Media (pp. 287-288).

(Publications Not part of thesis)

7. Jaglan, K., Pindiprolu, M. C., Sharma, T., Singam, A. R., **Goyal, N.**, Kumaraguru, P., & Brandes, U. (2024, May). Tight Sampling in Unbounded Networks. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 18, pp. 704-716).

8. **Goyal, N.**, Goel, A., Garg, T., Sachdeva, Kumaraguru, P. (2023, December). Efficient Knowledge Graph embeddings via Kernelized Random Projections. In Proceedings of Big Data Analytics in Astronomy, Science, and Engineering: 11th International Conference on Big Data Analytics, BDA 2023 (pp. 198-209).

Gratitude to PreCogers, IIITH, KRaCRs, and IIITD



Thanks to all Collaborators

- ❑ InfoEdge India Limited and Analytics team
- ❑ Dr. Niharika, Vijay, Rijula, Nitendra Rajput
- ❑ Anmol, Jushaan, Udit, Tanuj, Dr. Charu Sharma and Dr. Radhika Mamidi
- ❑ Thanks to Internal Committee members Dr. Rajiv Ratn Shah, Arun Balaji Buduru
- ❑ Microsoft India, ACM HT, InfoEdge India, CODS-COMAD for Conference Travel Grants

Acknowledgements



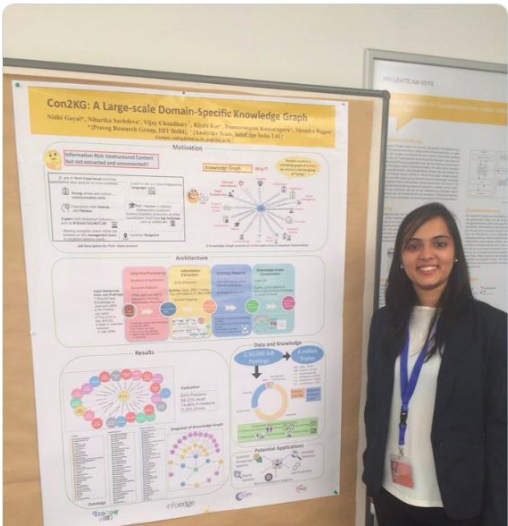
INDRAPRASTHA INSTITUTE of
INFORMATION TECHNOLOGY DELHI



Research Travels

ACM Hypertext 2019
@ACMHT

@nidhigoyalgoyal presenting her work "Con2KG: A Large-scale Domain-Specific Knowledge Graph" at our Poster Session #acmht19



GRW, 2023



RBCDSAI Web Science Symposium 2019, IIT Madras



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

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

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

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

Thank you
for your attention!

Definitions of High-quality data from Data.world and professional users

<https://data.world/blog/what-is-high-quality-data/>

Publications

5. **Goyal, N.**, Sachdeva, N., & Kumaraguru, P. (2021, August). Spy the lie: fraudulent jobs detection in recruitment domain using knowledge graphs. In Proceedings of Knowledge Science, Engineering and Management: 14th International Conference, KSEM 2021, Part II 14 (pp. 612-623). Springer International Publishing.

6. **Goyal, N.**, Sachdeva, N., Choudhary, V., Kar, R., Kumaraguru, P., & Rajput, N. (2019, September). Con2kg-a large-scale domain-specific knowledge graph. In Proceedings of the 30th ACM Conference on Hypertext and Social Media (pp. 287-288).

(Publications Not part of thesis)

7. Jaglan, K., Pindiprolu, M. C., Sharma, T., Singam, A. R., **Goyal, N.**, Kumaraguru, P., & Brandes, U. (2024, May). Tight Sampling in Unbounded Networks. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 18, pp. 704-716).

8. **Goyal, N.**, Goel, A., Garg, T., Sachdeva, Kumaraguru, P. (2023, December). Efficient Knowledge Graph embeddings via Kernelized Random Projections. In Proceedings of Big Data Analytics in Astronomy, Science, and Engineering: 11th International Conference on Big Data Analytics, BDA 2023 (pp. 198-209).

Publications

5. **Goyal, N.**, Sachdeva, N., & Kumaraguru, P. (2021, August). Spy the lie: fraudulent jobs detection in recruitment domain using knowledge graphs. In Proceedings of Knowledge Science, Engineering and Management: 14th International Conference, KSEM 2021, Part II 14 (pp. 612-623). Springer International Publishing.

6. **Goyal, N.**, Sachdeva, N., Choudhary, V., Kar, R., Kumaraguru, P., & Rajput, N. (2019, September). Con2kg-a large-scale domain-specific knowledge graph. In Proceedings of the 30th ACM Conference on Hypertext and Social Media (pp. 287-288).

(Publications Not part of thesis)

7. Jaglan, K., Pindiprolu, M. C., Sharma, T., Singam, A. R., **Goyal, N.**, Kumaraguru, P., & Brandes, U. (2024, May). Tight Sampling in Unbounded Networks. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 18, pp. 704-716).

8. **Goyal, N.**, Goel, A., Garg, T., Sachdeva, Kumaraguru, P. (2023, December). Efficient Knowledge Graph embeddings via Kernelized Random Projections. In Proceedings of Big Data Analytics in Astronomy, Science, and Engineering: 11th International Conference on Big Data Analytics, BDA 2023 (pp. 198-209).

Contribution : Improve quality of job postings



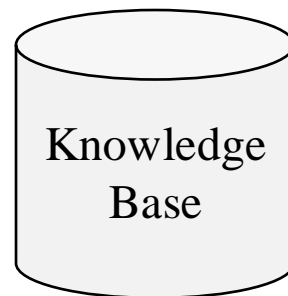
Unstructured text

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



OpenKB

This leads to redundant information and increased KB size.

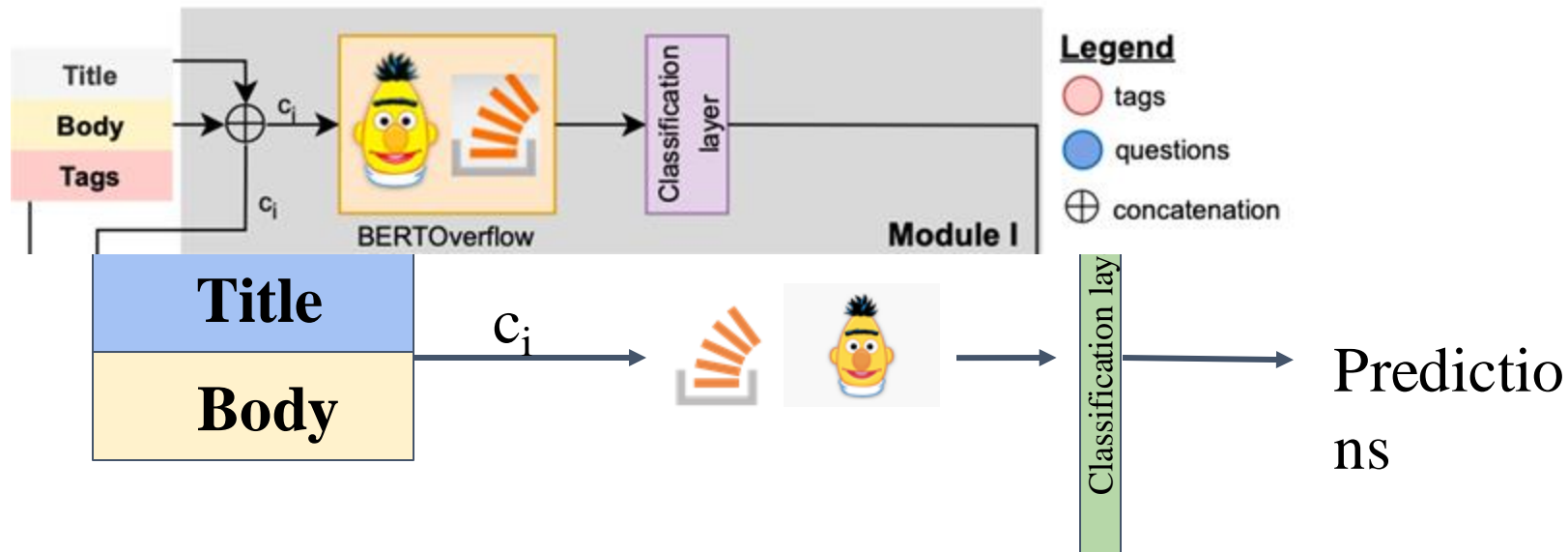


Performance

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

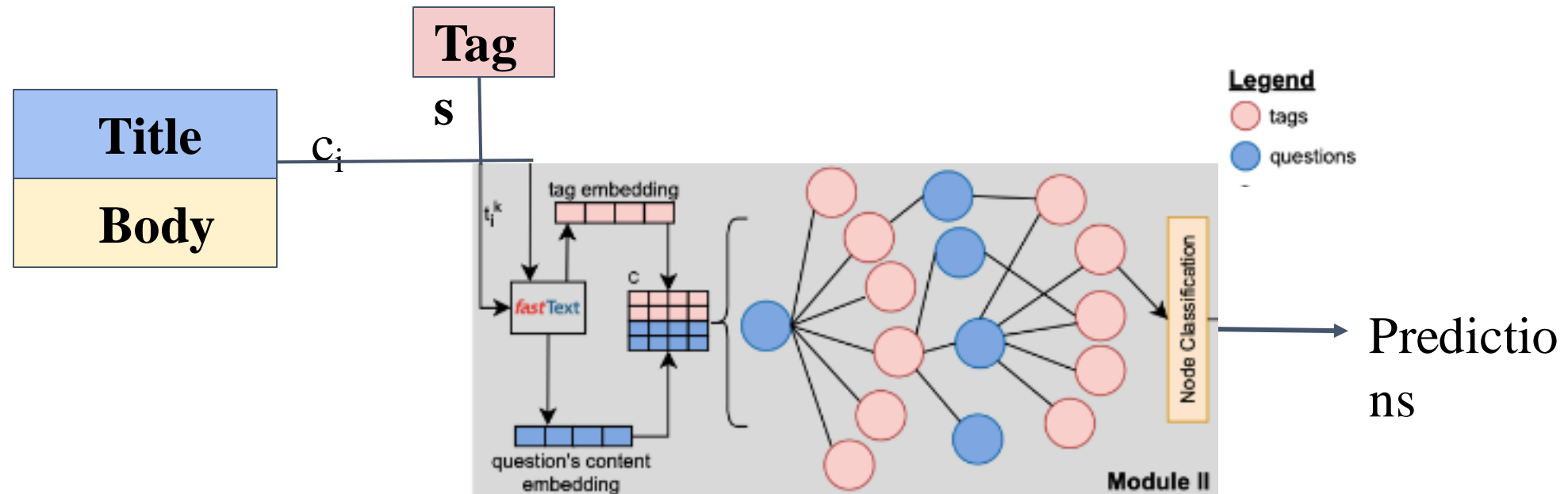


Proposed Approach



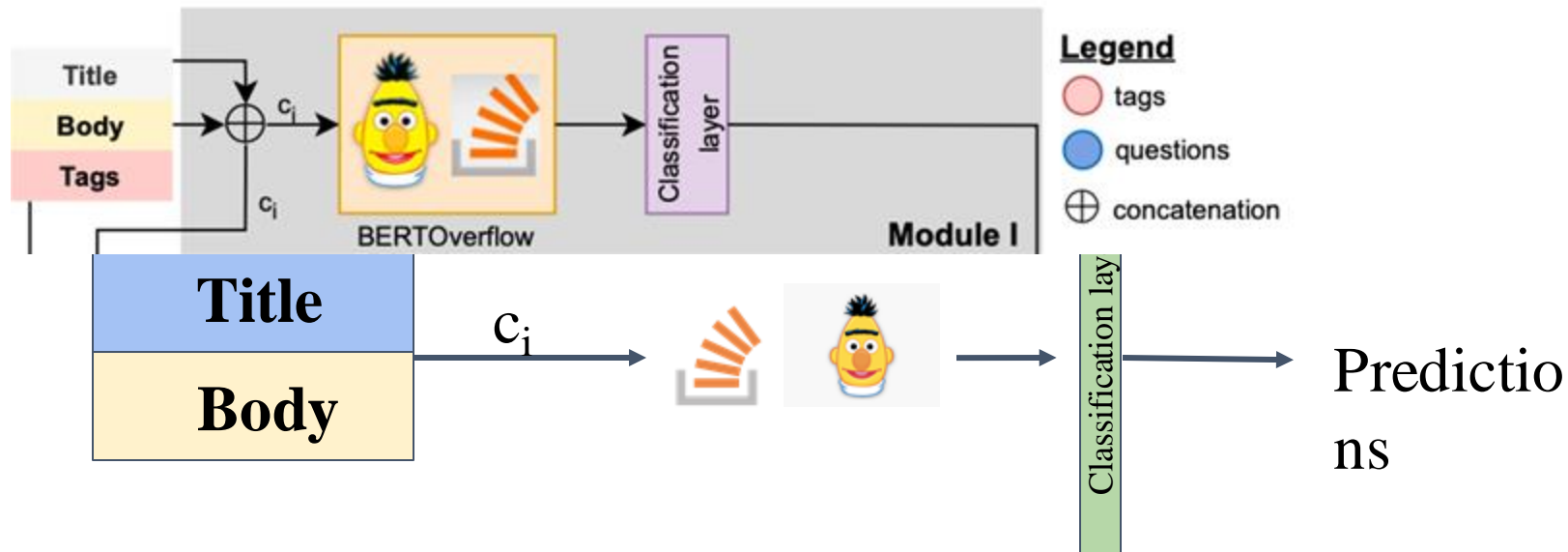
Module I fine-tunes the BERTOverflow model, which inputs the title and body for the question's classification task.

Proposed Approach



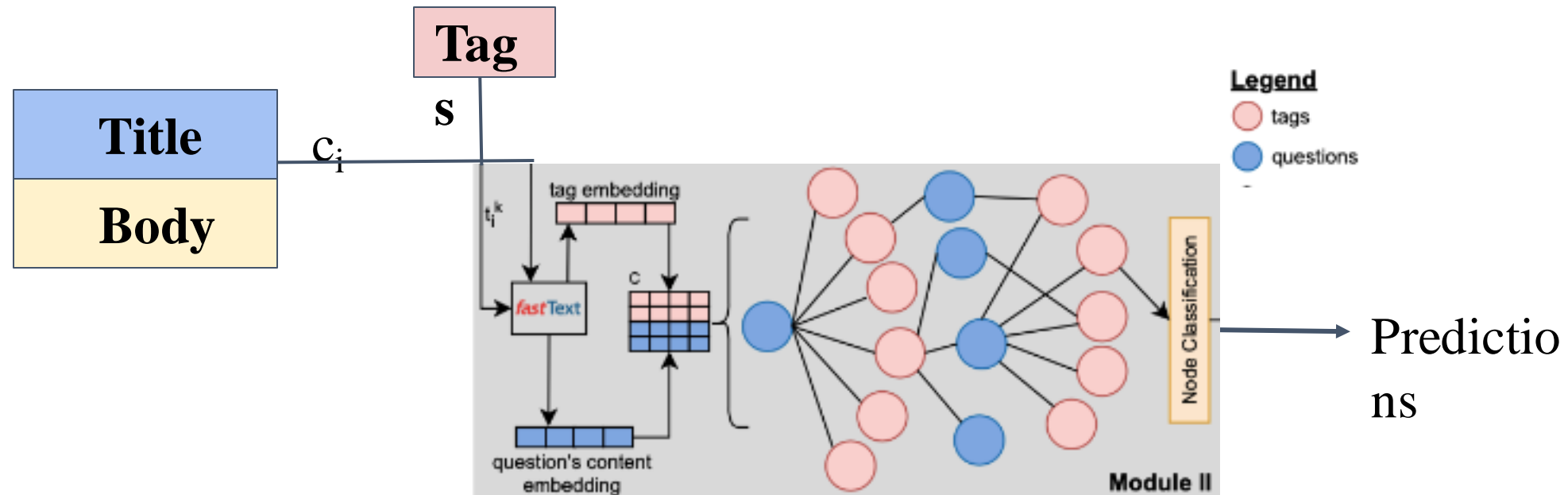
Module II

Proposed Approach



Module I fine-tunes the BERTOverflow model, which inputs the title and body for the question's classification task.

Proposed Approach



Module II

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>

Domain-specific knowledge graphs:

https://docs.google.com/presentation/d/1l2T8FlomxnP4jDC2mjMQLVJvhL2szd69D676GXGLqtA/edit#slide=id.ge1e1fc1e29_0_1316

Contributions 2

- Slide 23:

Functions :

https://docs.google.com/presentation/d/1_UZ1fpt4hZRPDvd0AAKaaiWf6Gq3-dcE60DPm2LTzig/edit#slide=id.ge4886c0253_0_50

Side Information Collection

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

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

Model	MRR		Hits @		
	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 search

Profiles from Search

Shannon Capper
Frontend Software Developer
Sammamish, Washington

Previous positions
Exercise Technician at G2 Science
Anatomy Lab Teaching Assistant

Education
California Polytechnic State University
Science (BS), Biology/Biological Sciences

178

Send InMail

Contact Info Edit

Recruiting Activity

All Activity Views (1)

Viewed by: Jensen Harris

New InMail message

To: Shannon Capper

Want to connect and chat?

No Salutation

I saw your profile on LinkedIn last week, and I was impressed by your **CREATIVITY** and **FEARLESS** approach. Did you see my message on Thursday?

We are Termfront and you can read a lot more about us on our website. We're looking for software engineers with your problem solving skills. We're growing fast and would love to have you on the team. You seem like a proven leader that knows what you're doing.

We're a small startup, but we're currently building the biggest problems in our industry, and we're looking for people who are FLEXIBLE about what days and times work best for them. I'm sure that you would like to work for a team that values how to reach you. It would be helpful to be able to reach you. Call me at 312-591-1234.

Please send a copy of your resume, and we'll be happy to review it. We're not currently looking for a career change, please let us know if you're interested.

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

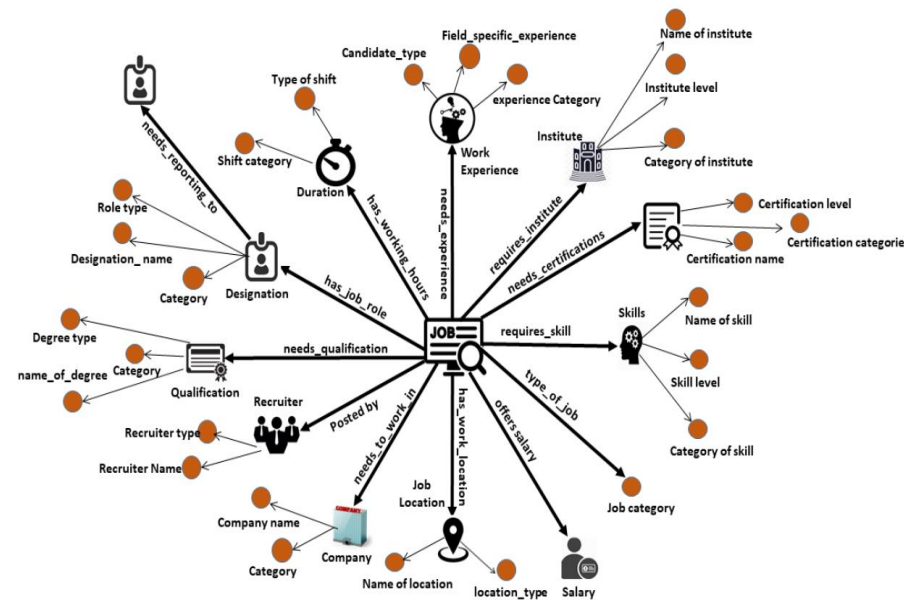
20s 30s 40s 50s 60s

- 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.



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}

¹Department of Informatics, Sokendai (The Graduate University for Advanced Studies)

²National Institute of Informatics, Tokyo, Japan

³National 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 Ngués 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.

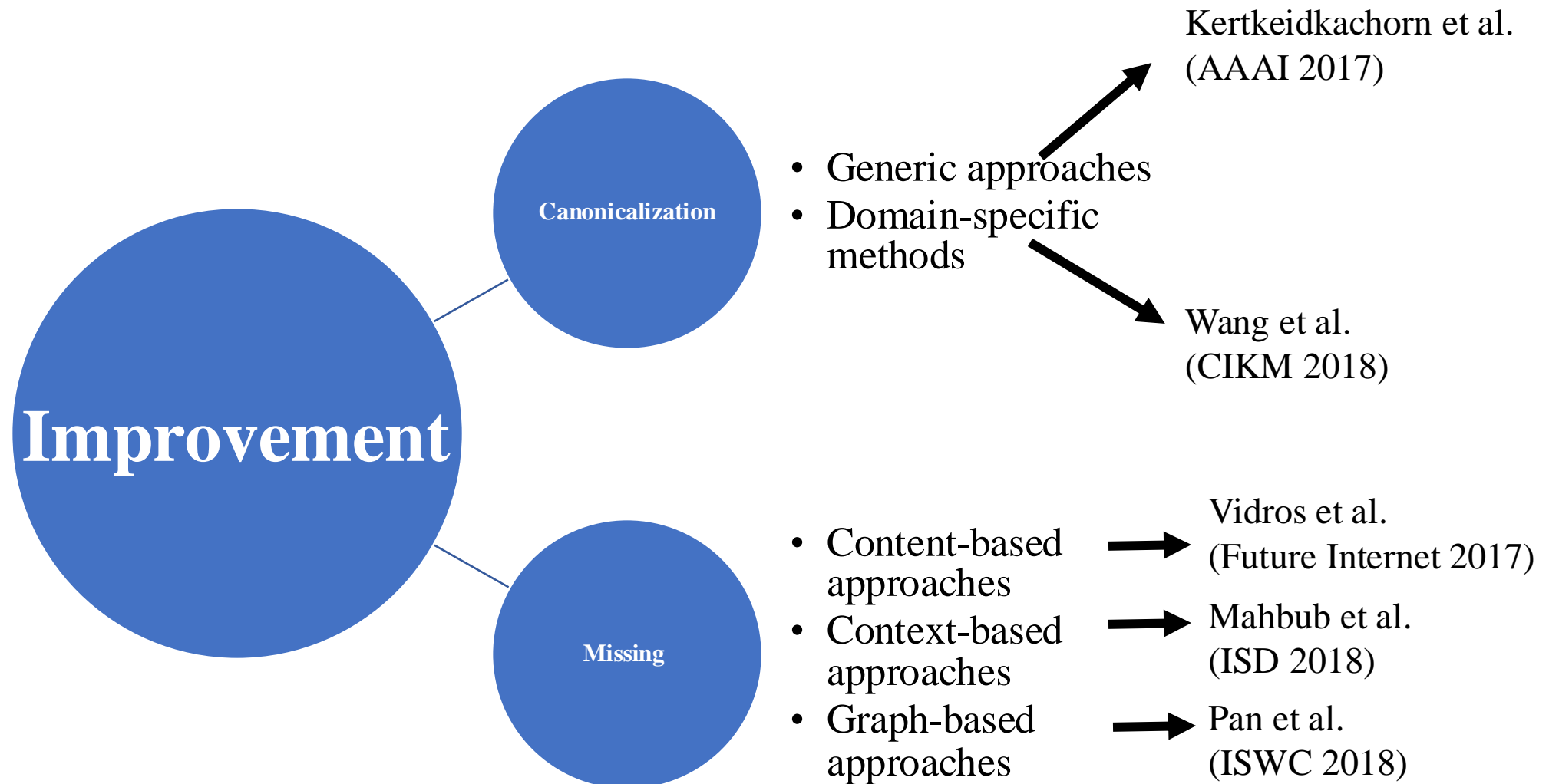
- (1) AceKG offers a heterogeneous academic information network, i.e., with multiple entity categories and relationship types, which supports researchers or engineers to conduct various academic data mining experiments.
- (2) AceKG is sufficiently large (3.13 billion triples with nearly 100G disk size)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

Literature Review



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’,
 {‘Names’,
 ‘title wikis’} - RDE(D)
 {‘Names’, ‘websites’, ‘title
 {‘Names’,
 ‘websites’, ‘affiliation’} - RDE(I)
 {‘Names’, ‘websites’, ‘title
 {‘Names’,
 wikis’, ‘types’} - ESCO(S)
 {‘types’, ‘industries’,
 ‘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.

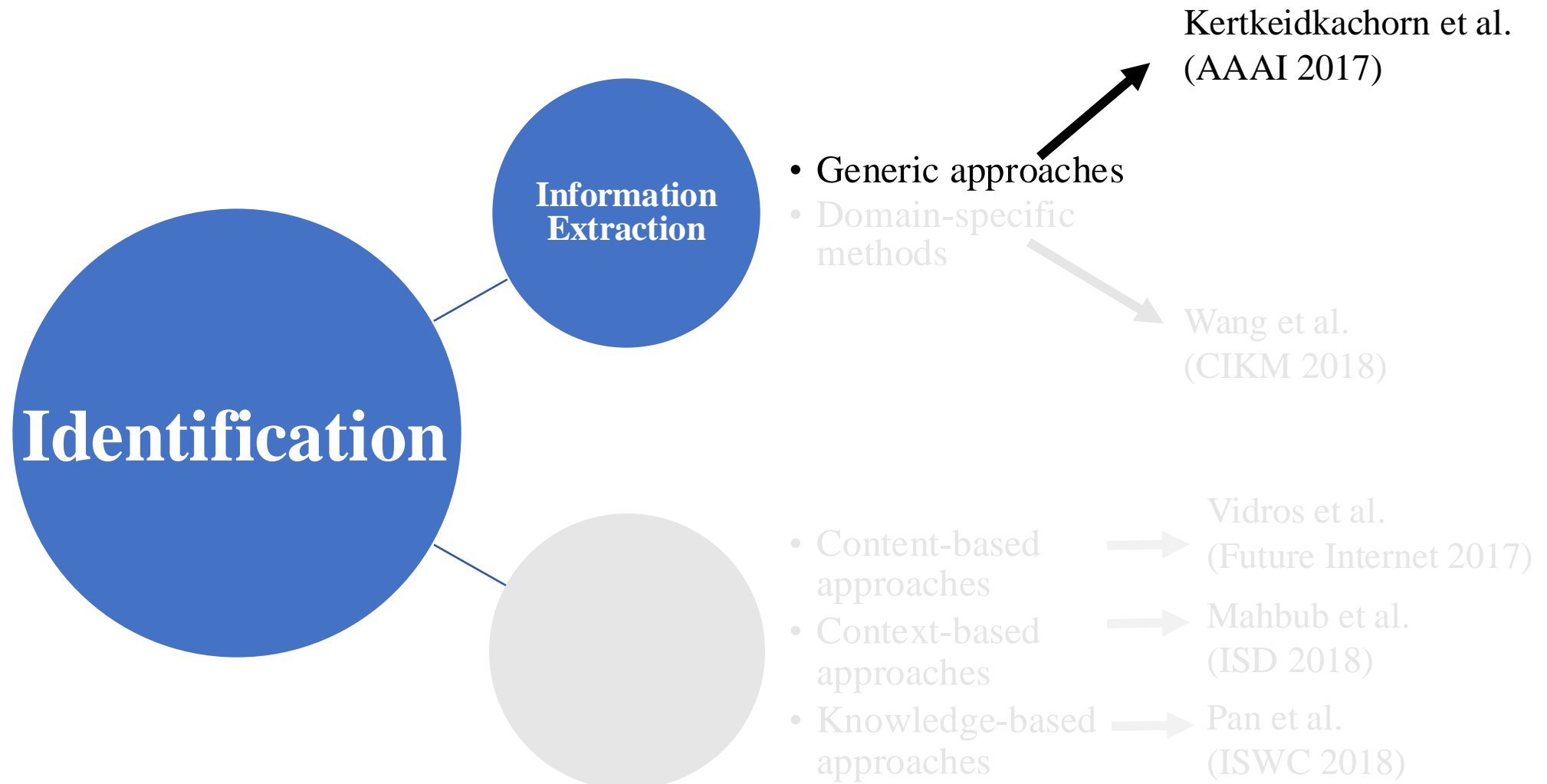
<p>Data Entry Clerks Position</p> <p>We have several openings available in this area earning \$1000.00-\$2500.00 per week. We are seeking only honest, self-motivated people with a desire to work in the home typing and data entry field, from the comfort of their own homes. The preferred applicants should 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.</p>	<p>Data Entry Clerk</p> <p>Responsibilities include, but are not limited to:</p> <ul style="list-style-type: none"> 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.

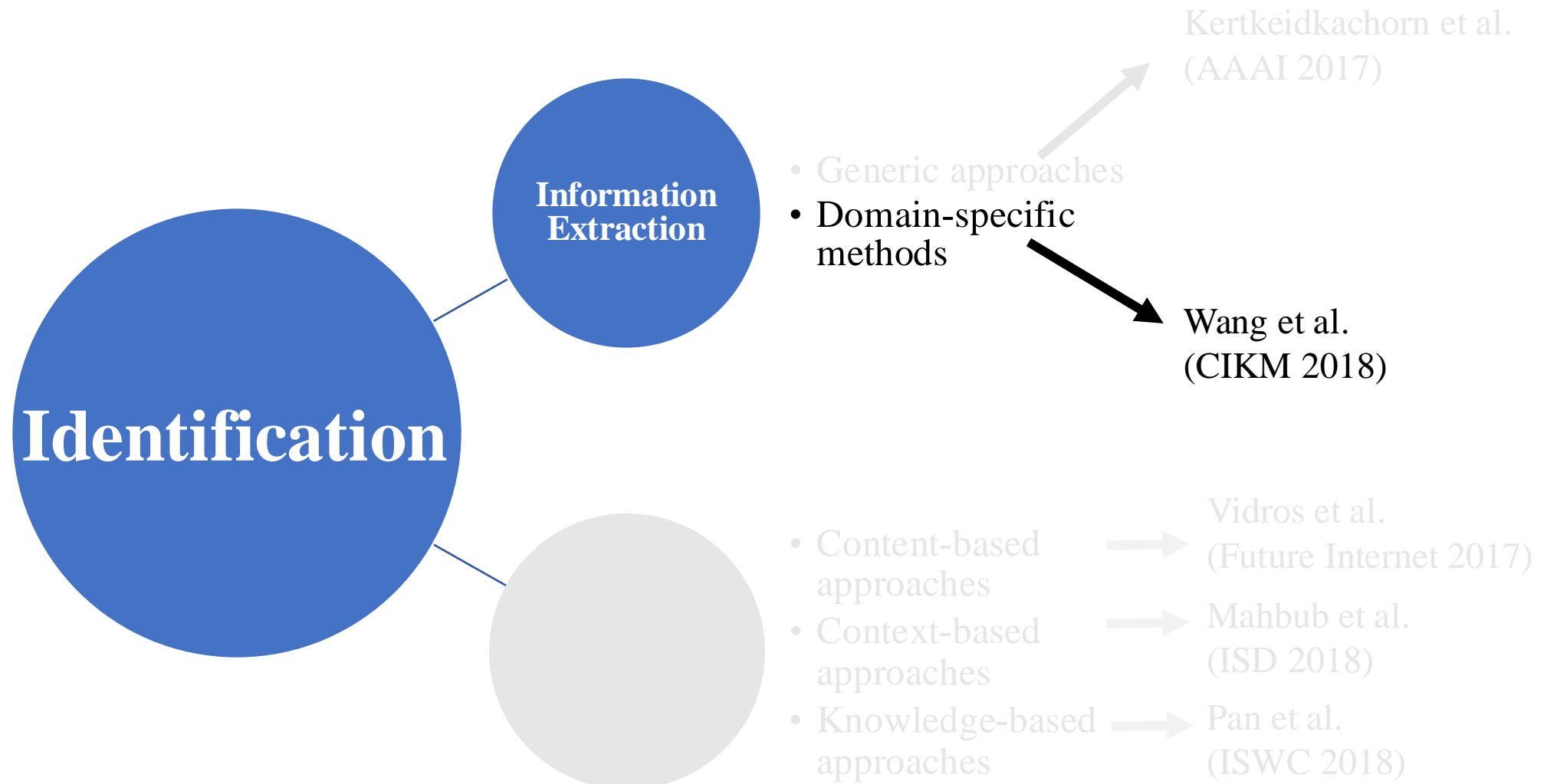
Literature Review

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

Literature Review



Literature Review



01

Standardize Entities

02

Find missing Entities

03

Identify misleading content

04

Identify low-quality content



Side Information Collection

We acquired additional knowledge from

Wikipedia InfoBox: Extracted knowledge from Wikipedia infoboxes for different datasets.

	{‘Names’, ‘websites’,
‘affiliation’} - RDE(I)	{‘Names’, ‘websites’,
‘title wikis’, ‘types’} - ESCO(S)	-
ESCO(D)	- 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.

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

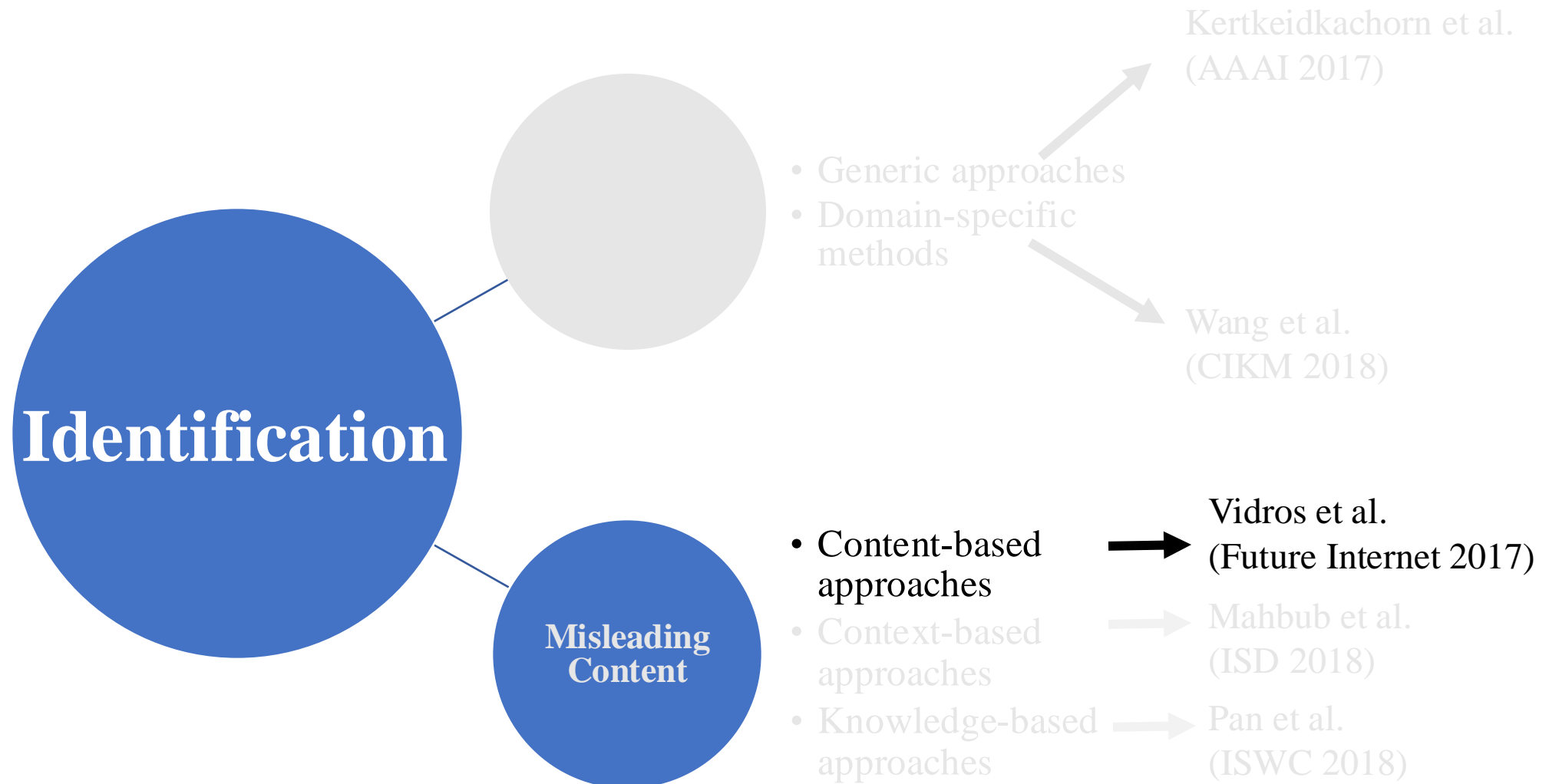
1. To **Identify** misleading content

- 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



Literature Review

Identi

Content Based Fake News Detection Using Knowledge Graphs

Jeff Z. Pan^{1(✉)}, Siyana Pavlova¹, Chenxi Li^{1,2}, Ningxi Li^{1,2}, Yangmei Li^{1,2}, and Jinshuo Liu^{2(✉)}

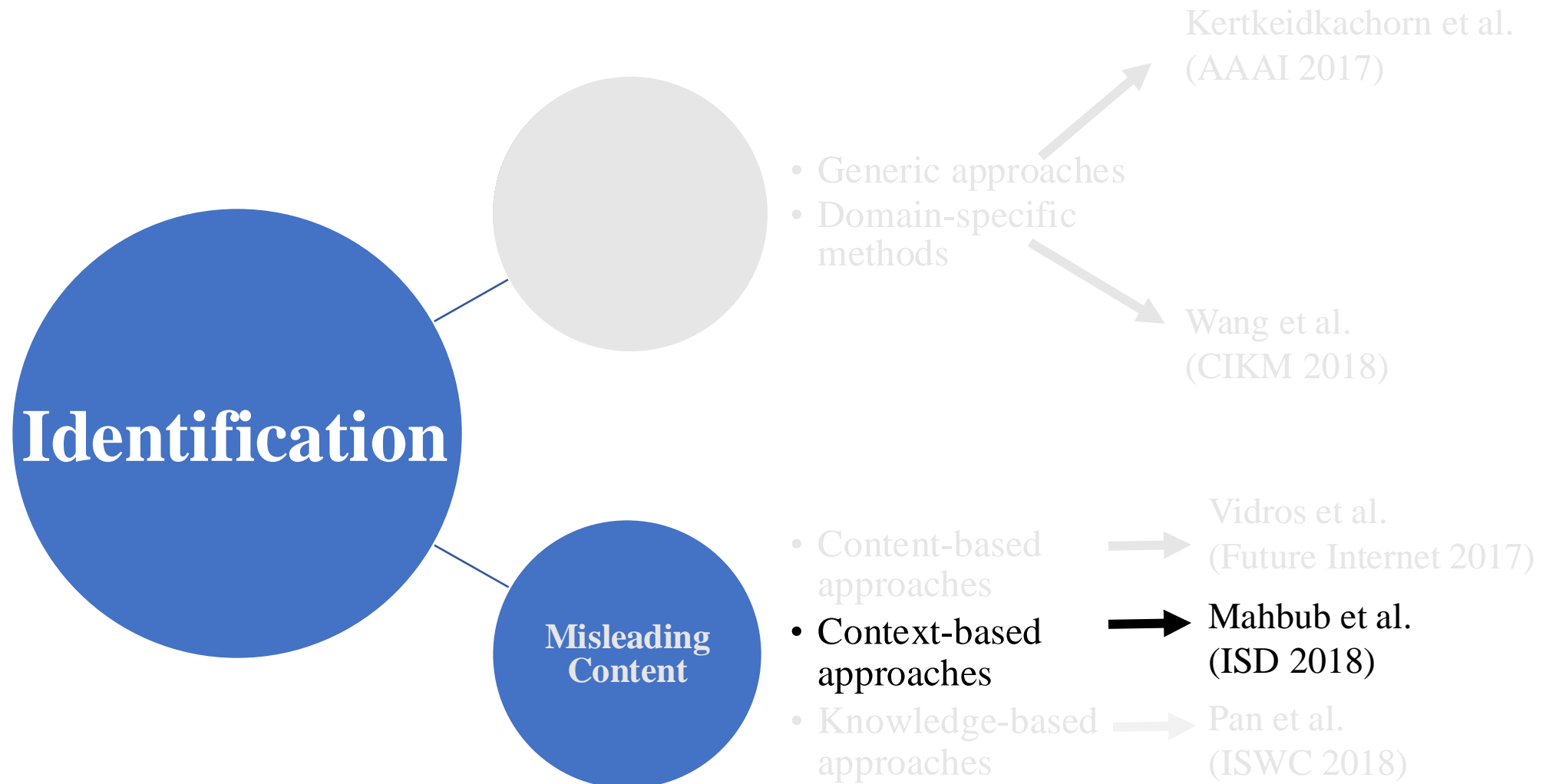
¹ University of Aberdeen, Aberdeen, UK
jeff.z.pan@abdn.ac.uk

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



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 Pavlova¹, Chenxi Li^{1,2}, Ningxi Li^{1,2}, Yangmei Li^{1,2}, and Jinshuo Liu^{2(✉)}

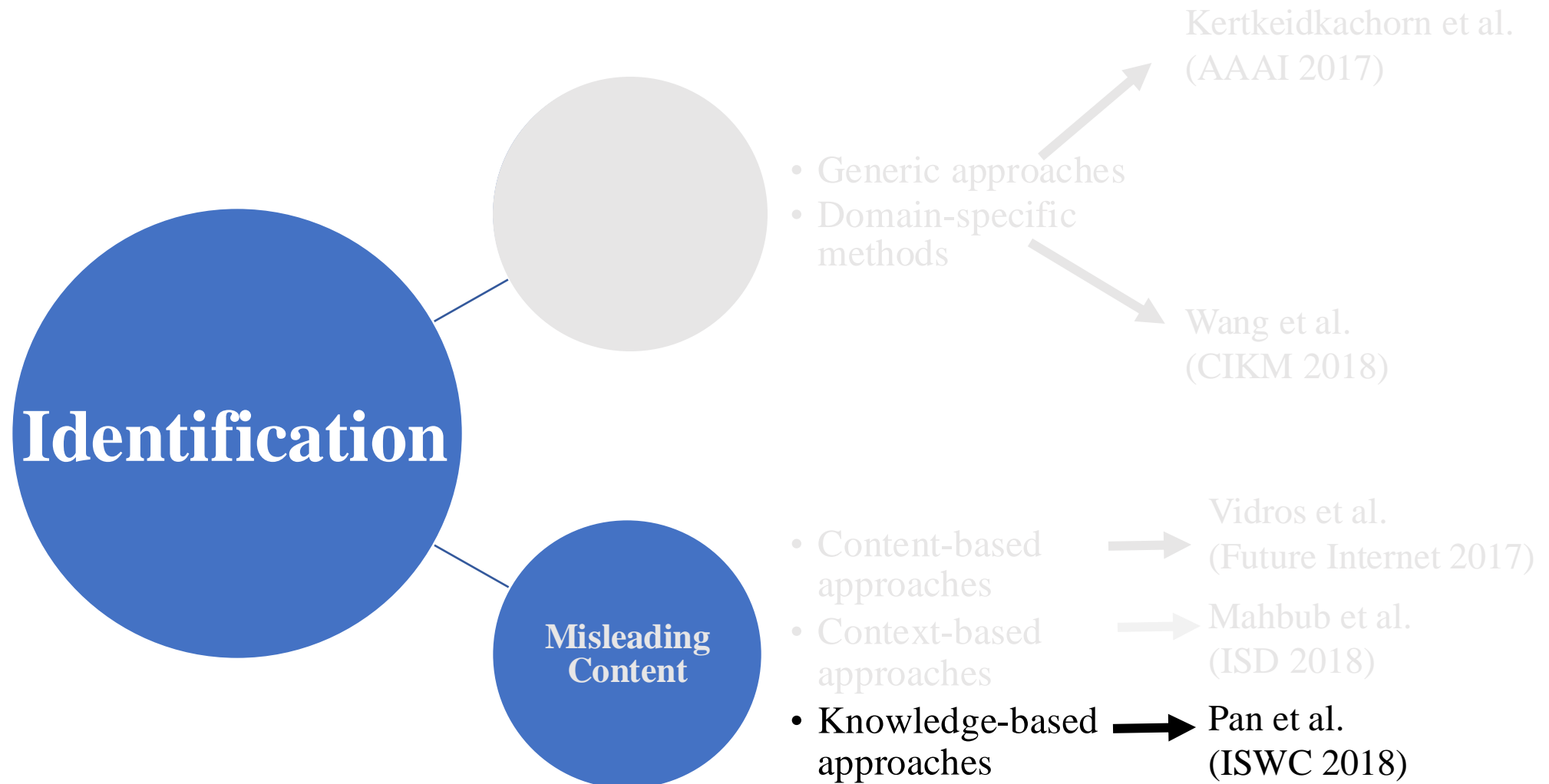
¹ University of Aberdeen, Aberdeen, UK
jeff.z.pan@abdn.ac.uk

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 Pavlova¹, 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

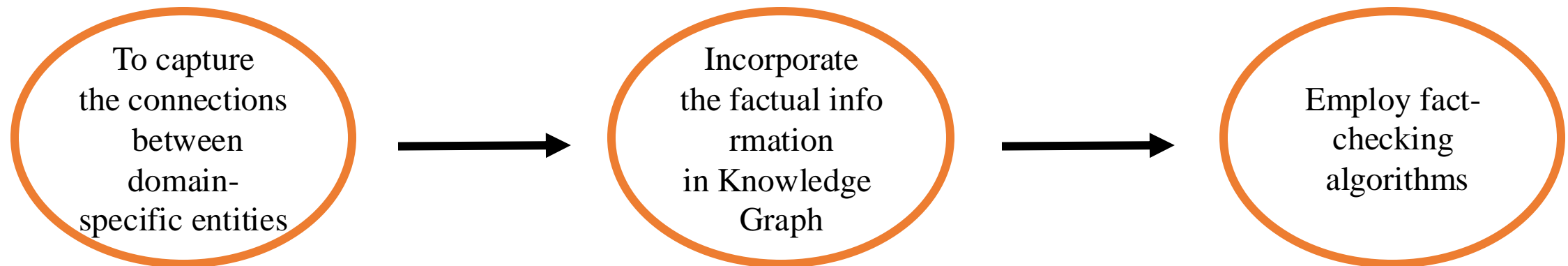
² Wuhan University, Wuhan, China
liujinshuo@whu.edu.cn

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
<p>Kertkeidkachorn et al. , T2KG: An End-to-End System for Creating Knowledge Graph (AAAI, 2017)</p> <p>Wang et al. (AceKG: A large-scale Knowledge Graph (CIKM, 2018)</p>	Domain-specific Knowledge Graphs	<ol style="list-style-type: none"> 1. Open (Public) Knowledge bases are available. They do not contain domain-specific information. 2. Recruitment domain-specific Knowledge bases are unavailable.
<p>Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset [4] (Future Internet, 2017)</p> <p>Content-based fake news Detection [3] (ISWC, 2020)</p>	Misleading	<ol style="list-style-type: none"> 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.

- 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 $J = \{J_1, J_2, J_3, \dots, J_N\}$ be the set of job postings and $Y = \{y_1, y_2, y_3, \dots, y_n\}$ be corresponding labels such that $y_i \in \{0, 1\}$. For every J_i , we extracted a set of triples T^i where $T^i = \{t^i_1, t^i_2, t^i_3, \dots, t^i_k\}$ 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

01

Standardize Entities

02

Find missing Entities

03

Identify misleading content

04

Identify low-quality content



Summary

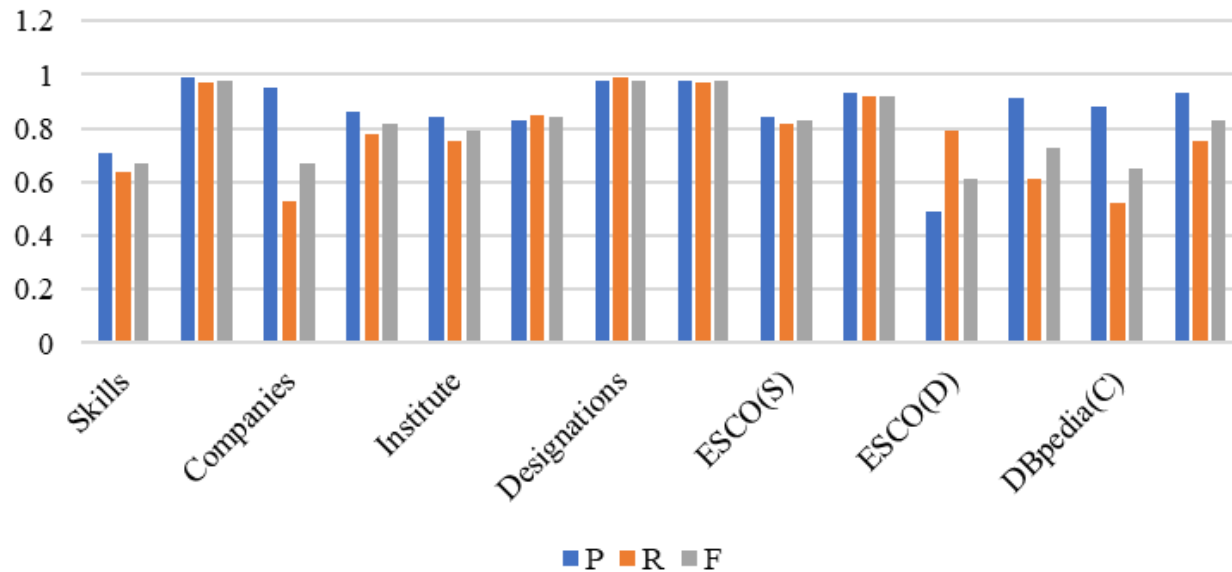
- We design a novel multi-tier framework Kernel-based Canonicalization Network (KCNet).
- KCNet induces a non-linear mapping between the contextual vector representations while capturing fine-granular and high-dimensional relationships among vectors.
- KCNet efficiently models more prosperous semantic 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 $\Phi: F(KG^A_{\text{false}}(T)^i, KG^A_{\text{true}}(T)^i, c^i, m^i)$ where $KG^A_{\text{true}}(T)^i$ is the scoring function, we learn from triple $t^i \in T^i | y_i = 0$ of legitimate job postings and $KG^A_{\text{false}}(T)^i$ 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.

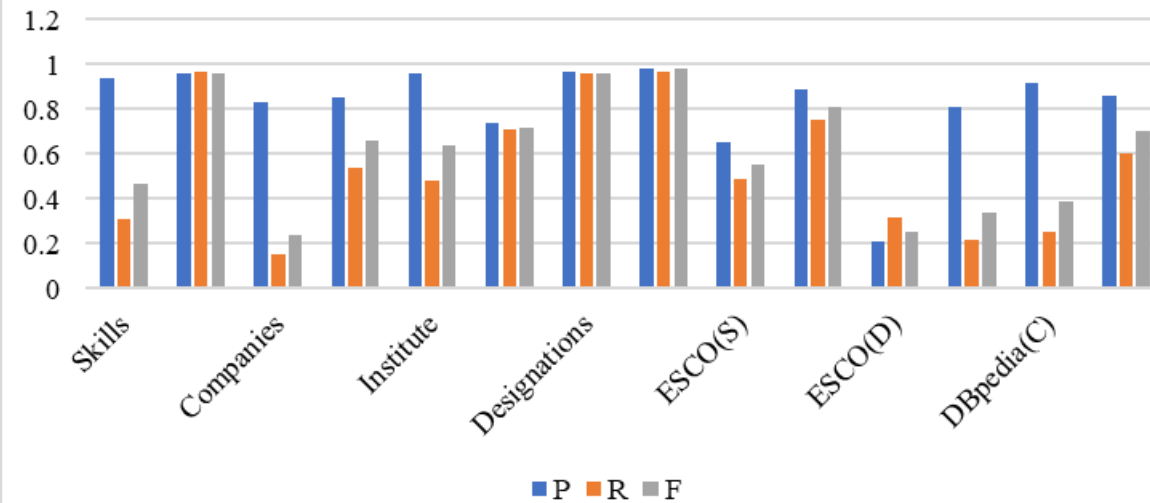
Test Results over HAC using pairwise similarity

Test Results over HAC using pairwise similarity



Micro

Test Results over HAC using pairwise similarity



Macro

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