ICANN21
30th International Conference on Artificial Neural Networks

KCNet: Kernel-based Canonicalization Network for entities in Recruitment Domain

Nidhi Goyal [1], Niharika Sachdeva [2], Anmol Goel [3], Jushaan Kalra [4], Ponnurangam Kumaraguru [5]

1. Indraprastha Institute of Information Technology, New Delhi, India
2. InfoEdge India Limited, Noida, India
3. Guru Gobind Singh Indraprastha University, Delhi, India
4. Delhi Technological University, Delhi, India
5. International Institute of Information Technology, Hyderabad, India

Canonicalization

Process of mapping multiple variations of a unique entity into the representative cluster

Microsoft corporation (india) pvt. ltd

University of Maryland, Baltimore

Microsoft corp india pvt ltd

umbc

Inspirisys

Microsoft, india

Accel Frontline

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

University of Maryland, Baltimore
umbc

Accel Frontline
Inspirisys

Motivation

1

Unstructured text

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

CV

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.

Challenges

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	University of Maryland, Baltimore umbc

Problem Formulation

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.

High-level overview of Approach

Clustering / Canonicalizing

3

Hierarchical Agglomerative clustering.

KCNet

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

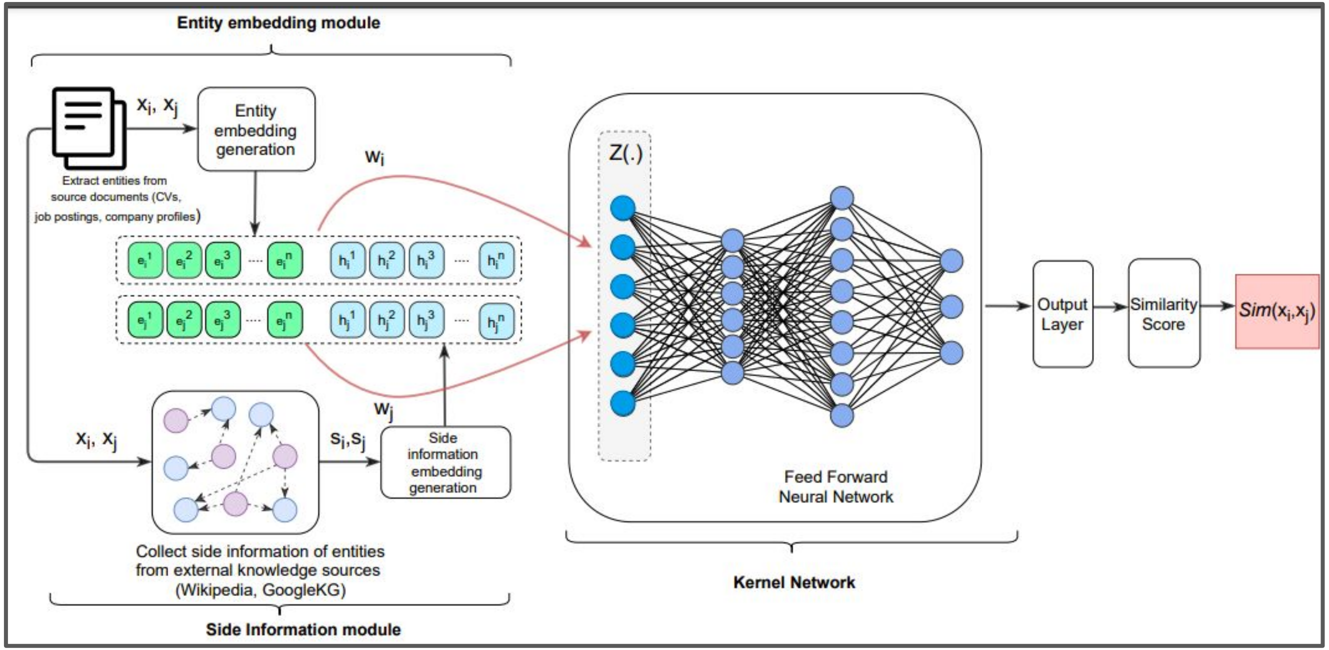
2

Dataset

1

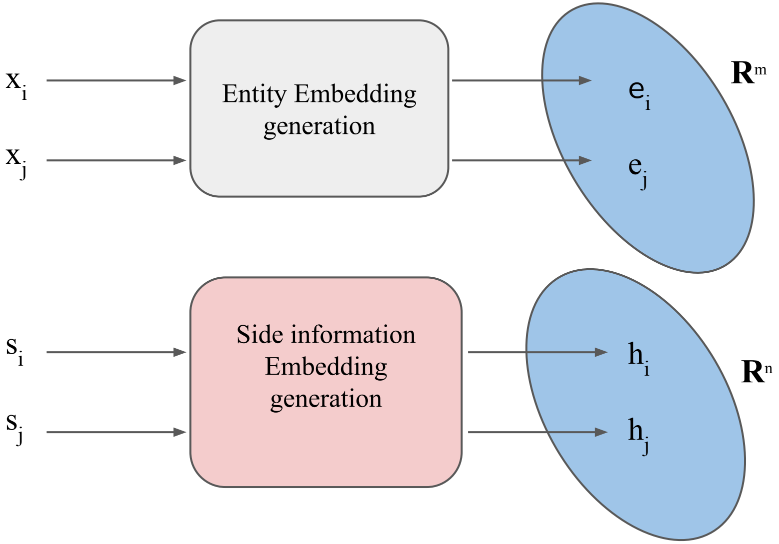
Entities extracted from documents.
Side Information Acquisition.
Positive and negative samples are created.
Random sampling is used for negative pairs.

Proposed Approach (KCNet)



Embedding Module

Given two entities $(\mathbf{x_i}, \mathbf{x_j})$ and their side information $(\mathbf{s_i}, \mathbf{s_j})$, embedding models produce $(\mathbf{e_i}, \mathbf{e_j}) \in \mathbf{R}^m$ and $(\mathbf{h_i}, \mathbf{h_j}) \in \mathbf{R^n}$



$\mathbf{w_i} = \mathbf{concat}\ (\mathbf{e_i}, \mathbf{h_i})$

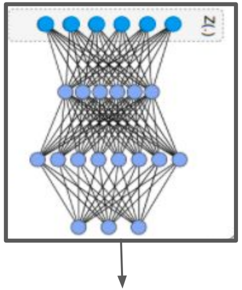
$\mathbf{w_j} = \mathbf{concat}\ (\mathbf{e_j}, \mathbf{h_j})$

Kernel Network

Z models element-wise relationships between input pairs.

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

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

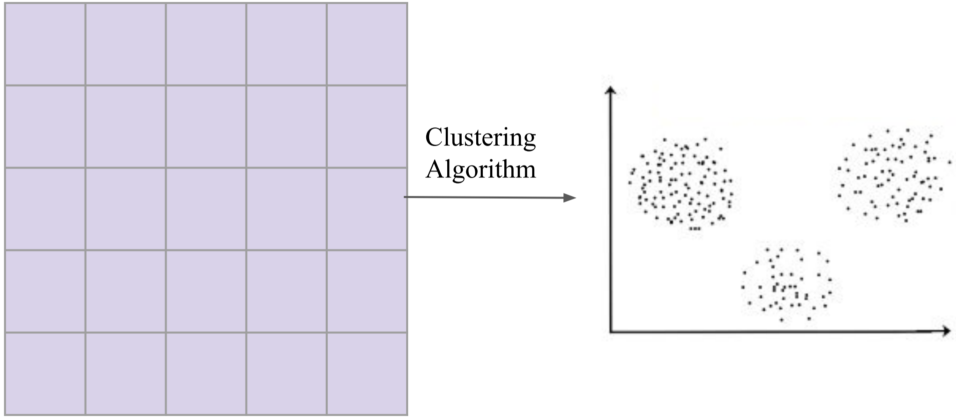


Similarity(x_i, x_j)

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

Clustering

- Clustering using pairwise similarity scores



Pairwise Distance Matrix

Canonical Clusters

Dataset Description

Source	Dataset	Entity Clusters
Proprietary	RDE(C)	25602
	RDE(I)	23690
	RDE(D)	3894
	RDE(S)	607
Open	DBpedia(C)	2944
	ESCO (S)	2644
	ESCO (D)	2903

Side information Collection

We acquired additional knowledge using:

1) 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).

2) 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.

Experiment Results

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 respectively. Results of † are taken from [4]. P and F refers to Precision and F1-scores. Distilled S-BERT (*, **) refers to (entity, entity $\odot$ side information) embedding using distilled S-BERT model.

Model	Performance													
	Proprietary								Open					
	S		D		I		C		ESCO(S)		ESCO(D)		DBpedia(C)	
	P	F	P	F	P	F	P	F	P	F	P	F	P	F
Galarraga-IDF†	33.2	12.5	63.0	60.3	64.3	66.5	75.8	71.2	50.8	32.8	61.7	38.9	22.6	23.6
Distilled S-BERT(*)+cosine	47.8	47.5	49.7	48.8	49.7	49.1	49.2	49.1	49.3	44.4	49.3	39.0	49.6	45.3
Distilled S-BERT(**)+cosine	47.5	48.8	49.8	49.9	34.6	41.5	56.2	48.4	49.5	50.0	49.4	49.7	50.0	49.8
CharBiLSTM+A†	81.8	86.9	72.6	77.2	84.5	84.8	99.3	98.9	85.9	86.9	76.3	75.1	72.1	59.7
WordBiLSTM+A†	80.1	86.5	90.5	94.8	80.6	83.3	95.3	95.6	85.6	89.6	83.1	83.7	77.6	70.7
CharBiLSTM+A+Word+A†	82.7	88.5	94.4	96.3	86.7	86.7	99.5	99.2	87.3	90.7	84.2	85.4	78.0	71.3
KCNet (without sideinfo)	**96.7**	**90.6**	**99.6**	90.9	**92.4**	**89.3**	99.4	98.8	**99.0**	**95.1**	**98.8**	**86.9**	**99.0**	**92.5**
KCNet (with sideinfo)	**99.5**	**99.4**	**99.7**	**99.6**	**99.5**	**99.5**	**99.5**	**99.3**	**99.2**	**98.3**	**98.8**	**89.4**	**99.1**	**97.0**

Experiment Results

- Char-BiLSTM+A+Word +A captures limited patterns and unable to model similar semantic variations (*mycology, fungi studies)* for which KCNet 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.

Table 2. Test Results over HAC using pairwise similarity. Here, β: baseline (*Char-BiLSTM+A+Word+A*) and γ: proposed model (*KCNet*) with sideinfo.

Dataset	Model	Metrics					
		Micro			Macro		
		P	R	F	P	R	F
S	β	0.71	0.64	0.67	0.94	0.31	0.47
	γ	**0.99**	**0.97**	**0.98**	**0.96**	**0.97**	**0.96**
D	β	0.95	0.53	0.67	0.83	0.15	0.24
	γ	0.86	**0.78**	**0.82**	**0.85**	**0.54**	**0.66**
I	β	0.84	0.75	0.79	0.96	0.48	0.64
	γ	0.83	**0.85**	**0.84**	0.74	**0.71**	**0.72**
C	β	0.98	0.99	0.98	0.97	0.96	0.96
	γ	0.98	0.97	**0.98**	**0.98**	**0.97**	**0.98**
ESCO(S)	β	0.84	0.82	0.83	0.65	0.49	0.55
	γ	**0.93**	**0.92**	**0.92**	**0.89**	**0.75**	**0.81**
ESCO(D)	β	0.49	0.79	0.61	0.21	0.32	0.25
	γ	**0.91**	0.61	**0.73**	**0.81**	0.22	**0.34**
DBpedia(C)	β	0.88	0.52	0.65	0.92	0.25	0.39
	γ	**0.93**	**0.75**	**0.83**	0.86	**0.60**	**0.70**

Conclusion

- 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.

Acknowledgements

Thank You

https://precog.iiit.ac.in