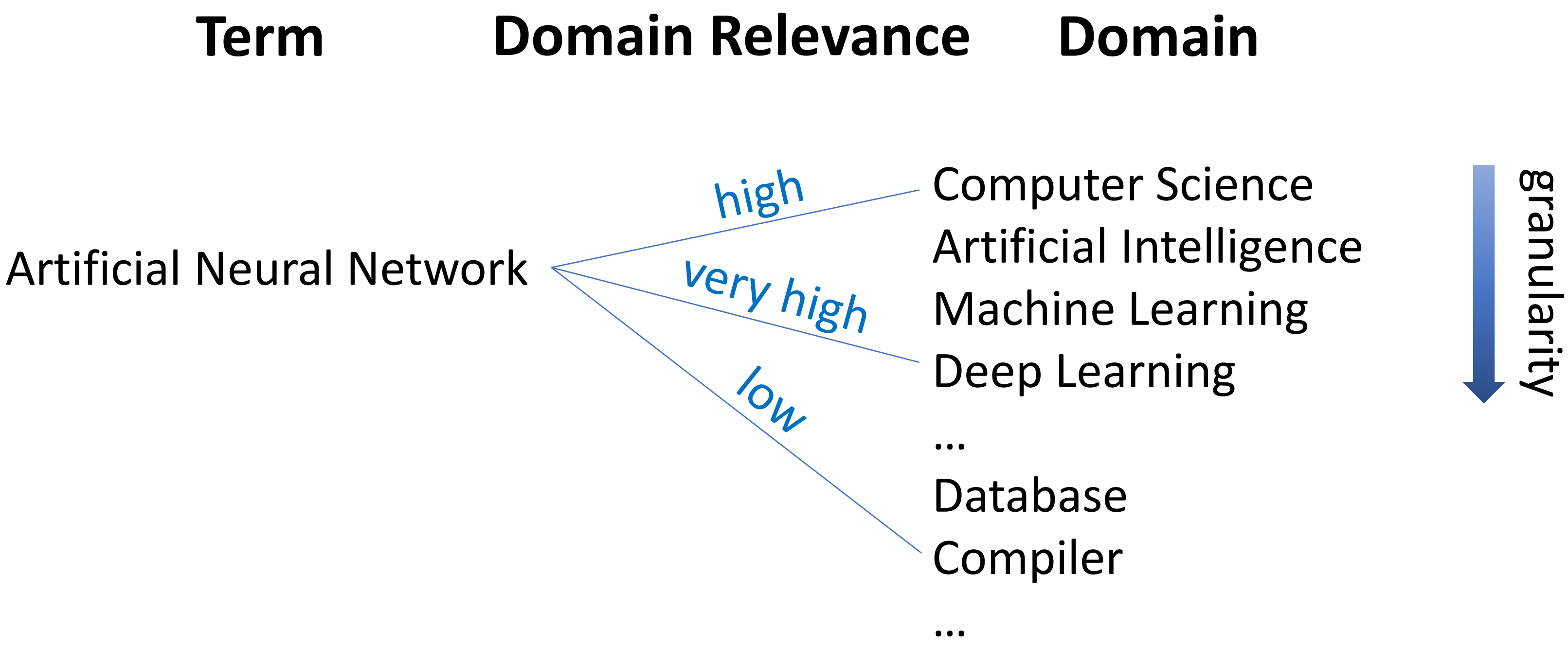
Measuring Fine-Grained Domain Relevance of Terms: A Hierarchical Core-Fringe Approach

Jie Huang, Kevin Chen-Chuan Chang, Jinjun Xiong, Wen-mei Hwu

University of Illinois at Urbana-Champaign
IBM Thomas J. Watson Research Center
IBM-Illinois Center for Cognitive Computing Systems Research (C3SR)

center for
cognitive computing
systems research

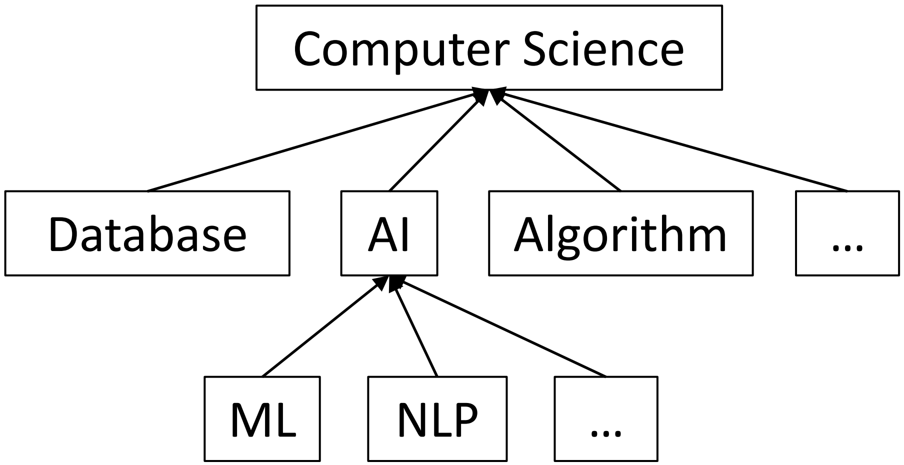
Fine-Grained Domain Relevance



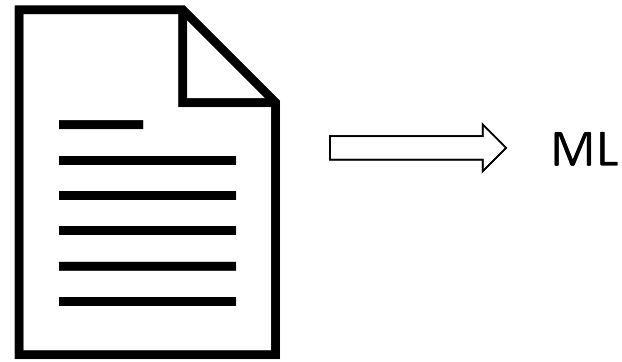
Fine-grained domain relevance:

The degree that a term is relevant to a given domain, and the given domain can be broad or narrow

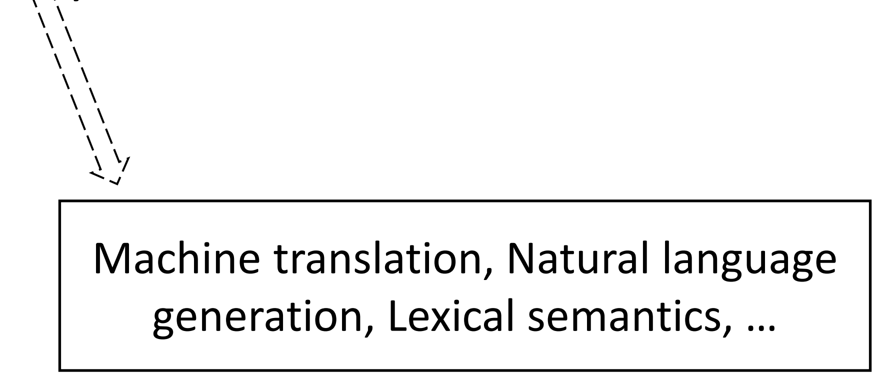
Applications



Domain taxonomy construction

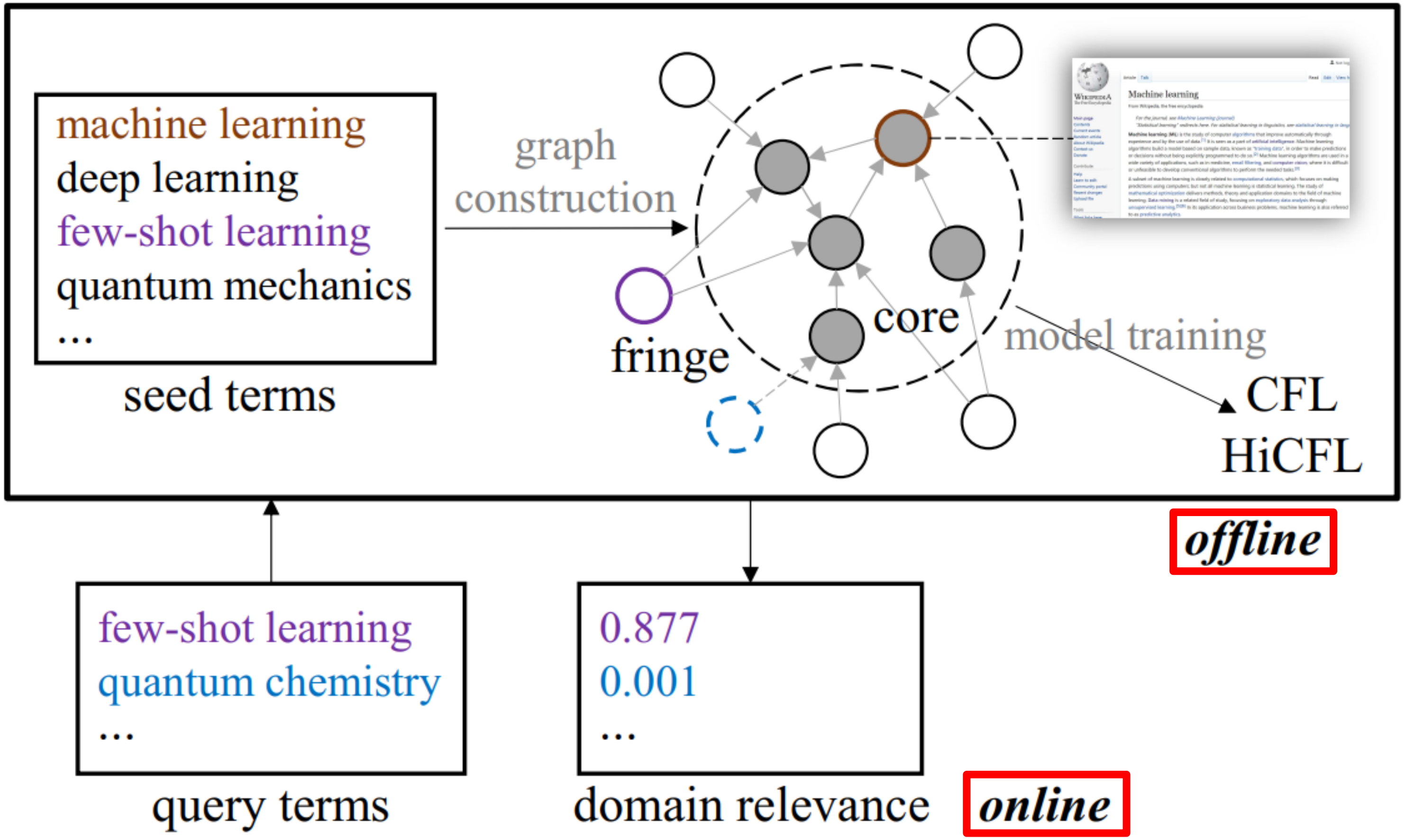


Document classification

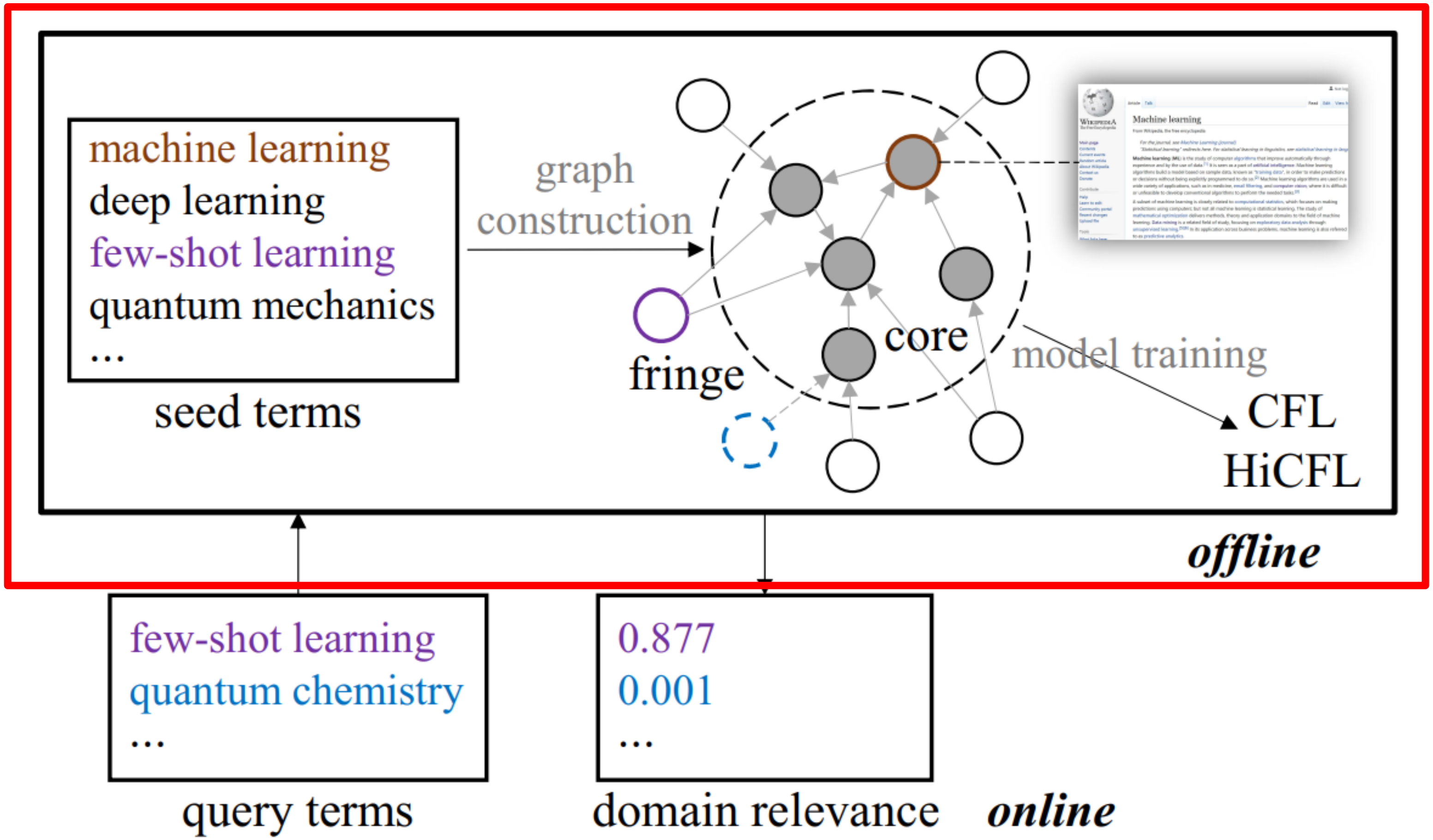


Domain-specific term extraction

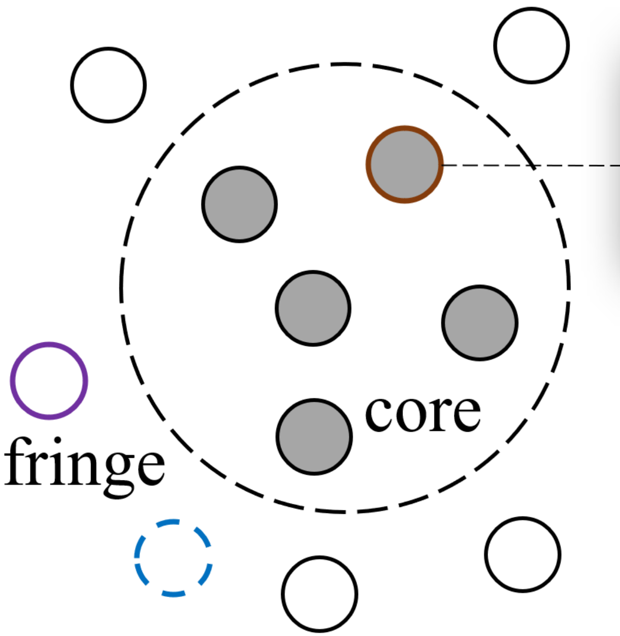
Method: Overview

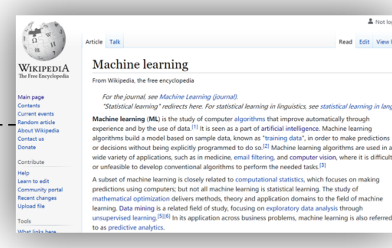


Method: Offline



Core-Anchored Semantic Graph



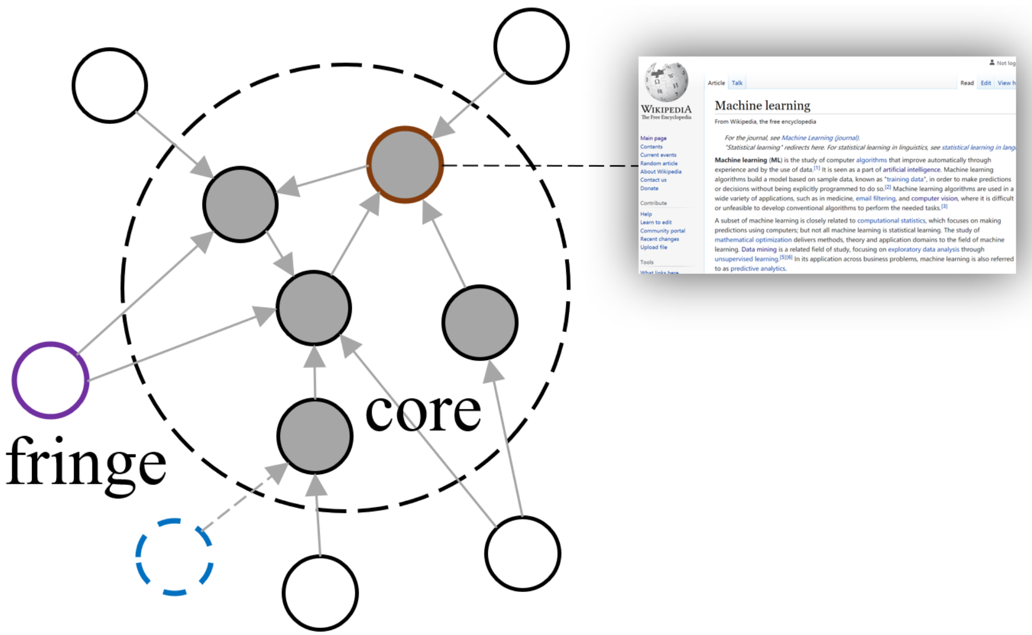


Core terms — domain relevance is easy to judge!

Core terms: Terms associated with rich description information (Wikipedia article pages), e.g., *Machine Learning*

Fringe terms: Terms without rich description information, e.g., *Few-Shot Learning* (usually the long-tail ones)

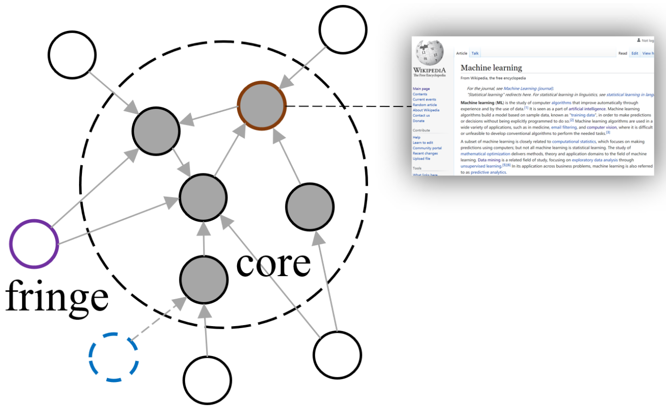
Core-Anchored Semantic Graph



Connect fringe terms to relevant core terms via document ranking

"Bridge domain relevance of terms through term relevance"

Core-Fringe Learning (CFL)



domain-relevant!

Assume labels of core terms (domain-relevant or not) are available:

$\Rightarrow$ ***a) propagate features of terms via term graph***

b) use labels of core terms for supervision

Graph convolution:

$$\boldsymbol{h}_i^{(l+1)} = \phi\Big(\sum_{j \in \mathcal{N}_i \cup \{i\}} \frac{1}{c_{ij}} \boldsymbol{W}_c^{(l)} \boldsymbol{h}_j^{(l)} + \boldsymbol{b}_c^{(l)} \Big)$$

Loss:

$$\mathcal{L} = - \sum_{i \in \mathcal{V}_{core}} (y_i \log \boxed{z_i} + (1 - y_i) \log(1 - z_i))$$

Domain Relevance

Hierarchical Core-Fringe Learning (HiCFL)

Hierarchy of domains: CS -> AI -> ML

"An ML *term should also be relevant to* CS*"* => ***Hierarchical Learning***

Global information:

$$\boldsymbol{z}_p = \sigma(\boldsymbol{W}_p^{(l_p)} \boldsymbol{a}_p^{(l_p)} + \boldsymbol{b}_p^{(l_p)})$$

$$\boldsymbol{a}_p^{(l+1)} = \phi(\boldsymbol{W}_p^{(l)} [\boldsymbol{a}_p^{(l)}; \boldsymbol{h}^{(l_c)}] + \boldsymbol{b}_p^{(l)})$$

Local information:

$$\boldsymbol{z}_q^{(l)} = \sigma(\boldsymbol{W}_q^{(l)} \boldsymbol{a}_q^{(l)} + \boldsymbol{b}_q^{(l)})$$

$$\boldsymbol{a}_q^{(l)} = \phi(\boldsymbol{W}_t^{(l)} \boldsymbol{a}_p^{(l)} + \boldsymbol{b}_t^{(l)})$$

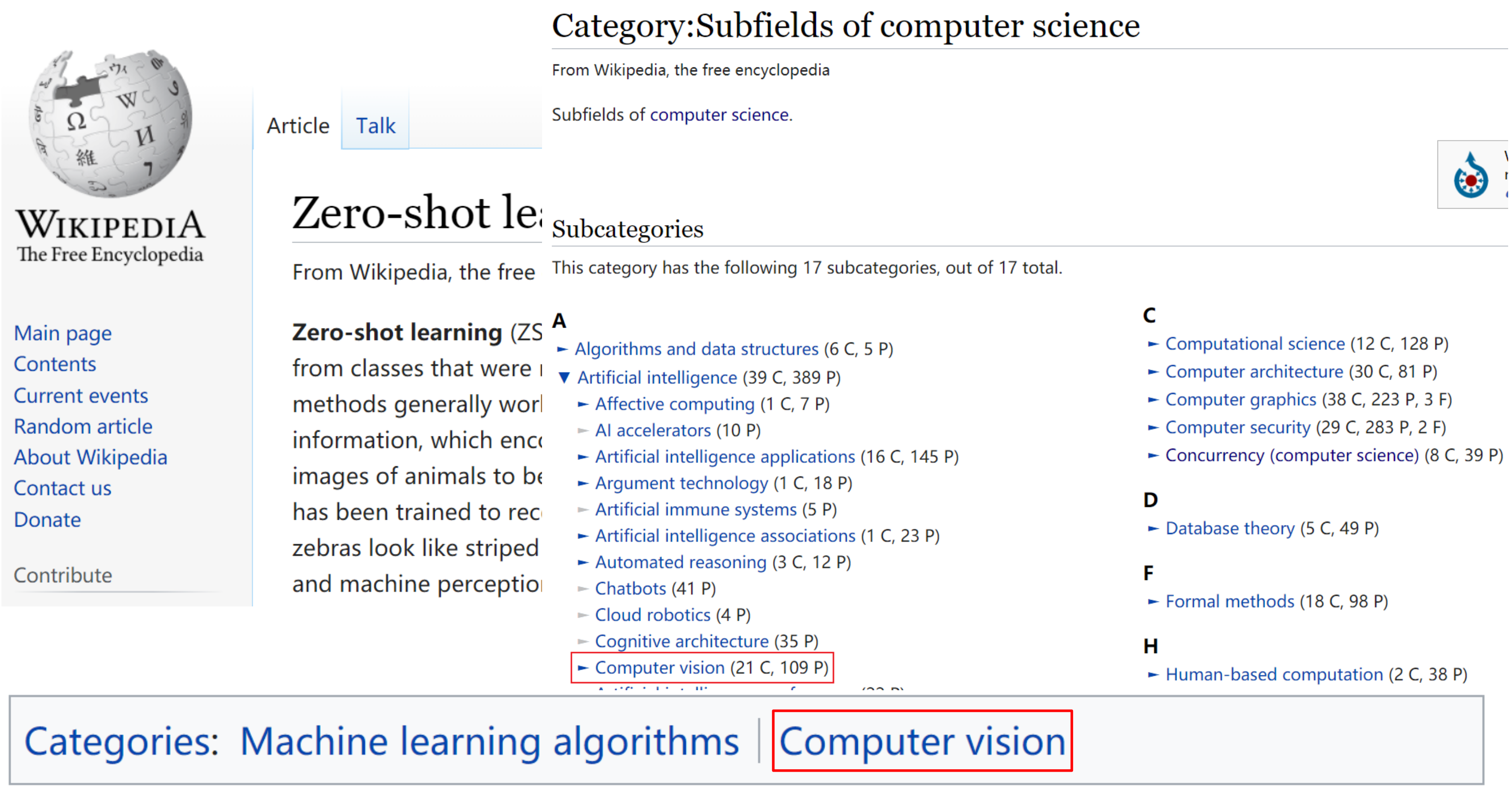
Loss:

$$\mathcal{L}_h = \epsilon(\boldsymbol{z}_p, \boldsymbol{y}^{(l_p)}) + \sum_{l=1}^{l_p} \epsilon(\boldsymbol{z}_q^{(l)}, \boldsymbol{y}^{(l)})$$

Domain relevance:

$$\boldsymbol{s} = \alpha \cdot \boldsymbol{z}_p + (1 - \alpha) \cdot (\boldsymbol{z}_q^{(1)} \circ \boldsymbol{z}_q^{(2)}, \ldots, \boldsymbol{z}_q^{(l_p)})$$

Automatic Annotation



Machine learning algorithms and *Computer vision* are categories in CS

=> ***Zero-shot learning is domain-relevant!***

Hierarchical Positive-Unlabeled Learning

For narrow domains like *Deep Learning*, a category tree might not be available in Wikipedia. ***Automatic Annotation is impractical***

$\Rightarrow$ **Hierarchical Positive-Unlabeled (PU) Learning**

For *Deep Learning* (DL),

Hierarchy of domains: $\underbrace{\text{CS -> AI -> ML}}$ -> DL

Automatic Annotation

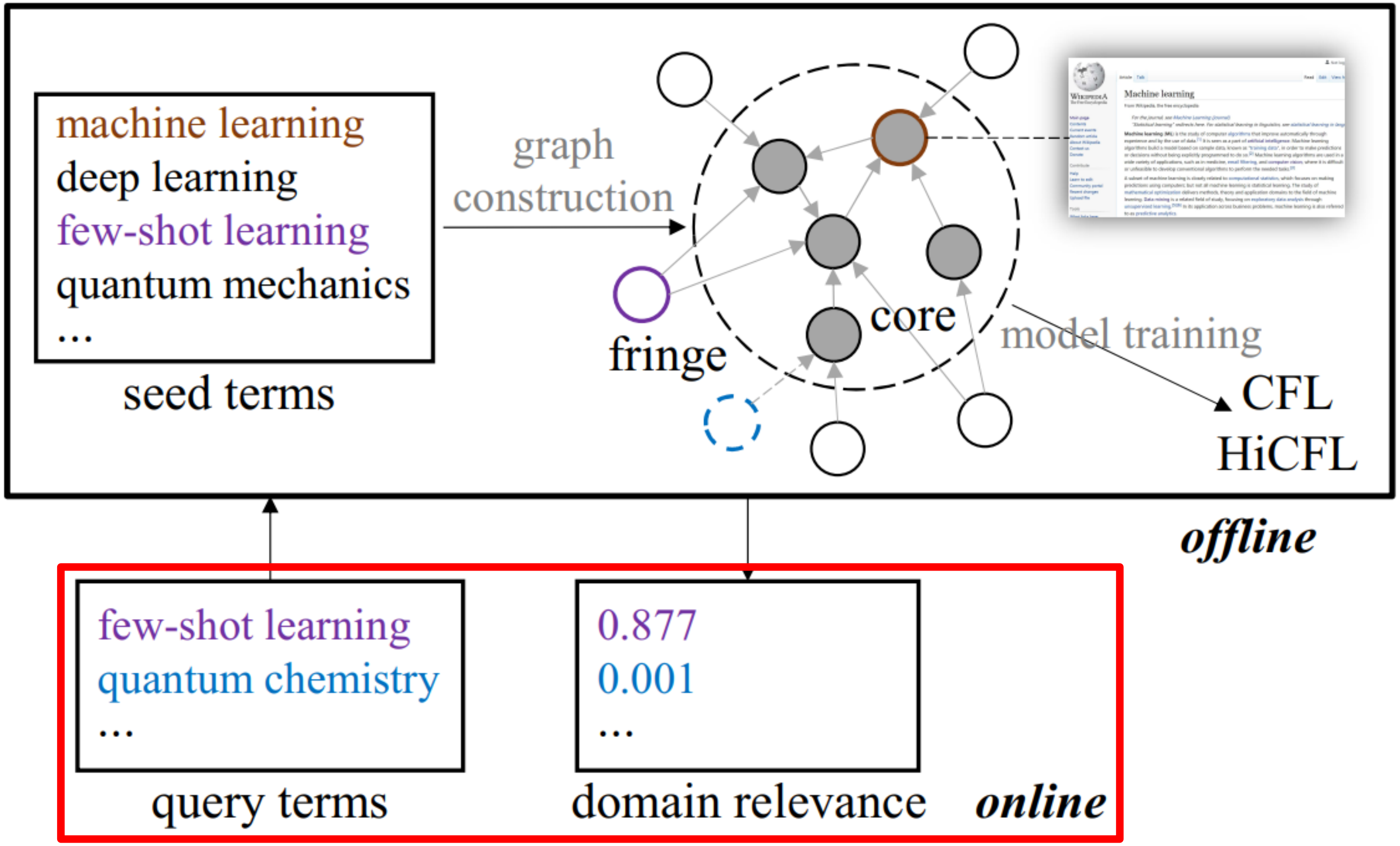
Positive: *k* DL terms provided by users

Negative: non-ML terms

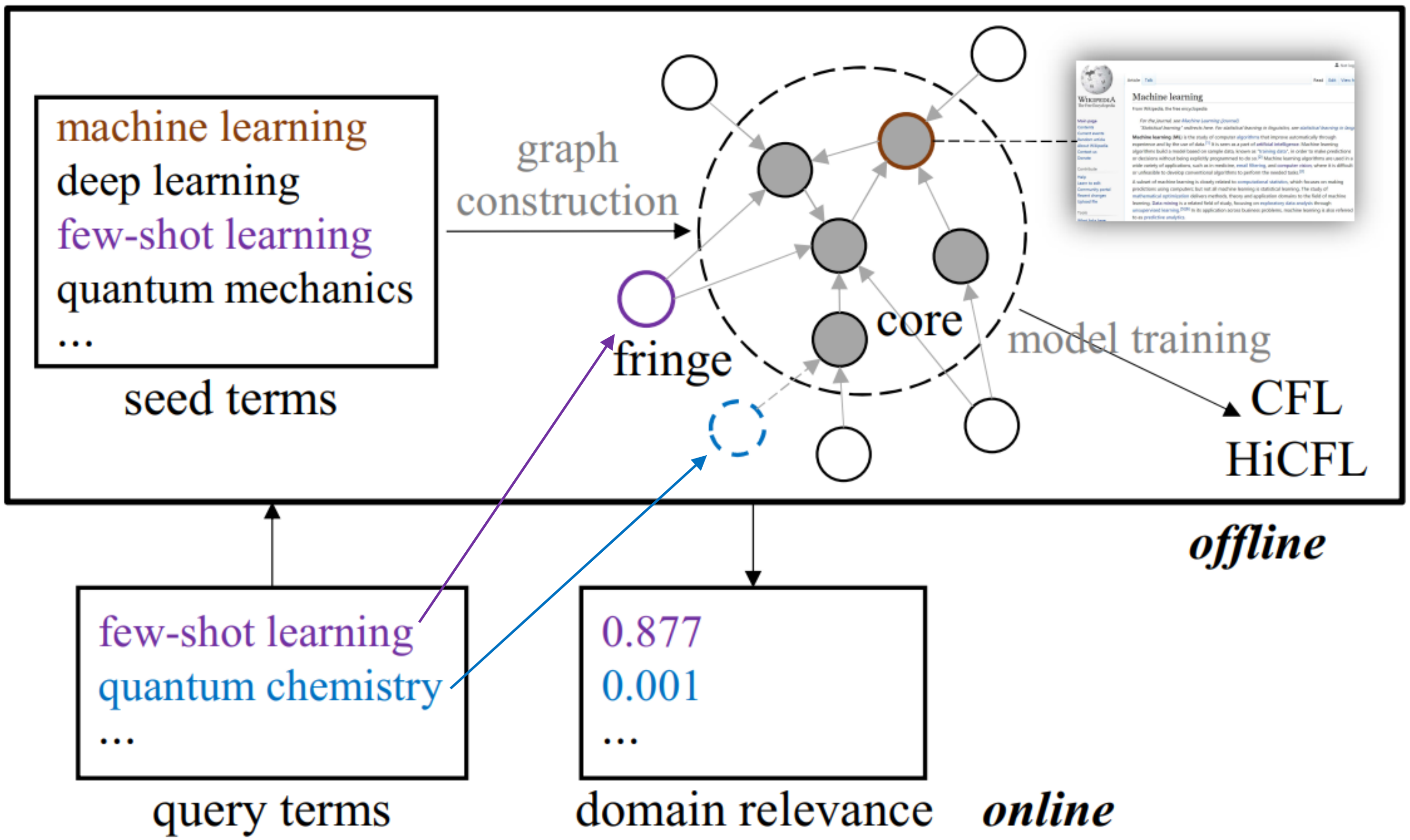
HiCFL

$$\mathcal{L}_h = \epsilon(\boldsymbol{z}_p, \boldsymbol{y}^{(l_p)}) + \sum_{l=1}^{l_p} \epsilon(\boldsymbol{z}_q^{(l)}, \boldsymbol{y}^{(l)})$$

Method: Online



Method: Online



Experiments: Overview

Comparison to Baselines: Compare with existing methods on *Automatic Term Extraction*

Comparison to Human Performance: Compare with human professionals

Case Studies: Case studies for ML and DL

domain		#terms	core ratio
CS	ML	113,038	27.7%
Phy	QM	416,431	12.1%
Math	AA	103,984	26.4%

Statistics of data. 3 broad domains, 3 narrow domains

Comparison to Baselines

		Computer Science		Physics		Mathematics	
		ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
RDF	SG	0.714	0.417	0.736	0.496	0.694	0.579
LR	G	0.802±0.000	0.535±0.000	0.822±0.000	0.670±0.000	0.854±0.000	0.769±0.000
MLP	S	0.819±0.003	0.594±0.003	0.853±0.001	0.739±0.004	0.868±0.000	0.803±0.001
MLP	G	0.863±0.001	0.674±0.002	0.874±0.001	0.761±0.003	0.904±0.001	0.846±0.002
MLP	SG	0.867±0.001	0.667±0.002	0.875±0.001	0.765±0.002	0.904±0.001	0.843±0.003
MC	SG	0.868±0.002	0.664±0.006	0.877±0.003	0.768±0.004	0.903±0.001	0.843±0.002
CFL	G	**0.885**±0.001	**0.712**±0.002	**0.905**±0.000	**0.812**±0.002	0.918±0.001	0.870±0.002
CFL	C	0.883±0.001	0.708±0.002	0.901±0.000	0.800±0.001	**0.919**±0.001	**0.879**±0.002

S and G indicate the corpus used. S: domain-specific corpus, G: general corpus, SG: both.
C means the pre-trained compositional GloVe embeddings are used.

CFL outperforms baselines significantly

⇒ *Core-anchored semantic graph and feature aggregation are helpful!*

⇒ *Domain relevance can be bridged via term relevance!*

Comparison to Baselines

		Machine Learning		Quantum Mechanics		Abstract Algebra	
		ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
LR	G	0.917$\pm$0.000	0.346$\pm$0.000	0.879$\pm$0.000	0.421$\pm$0.000	0.872$\pm$0.000	0.525$\pm$0.000
MLP	S	0.902$\pm$0.001	0.453$\pm$0.009	0.903$\pm$0.001	0.545$\pm$0.004	0.910$\pm$0.000	0.641$\pm$0.007
MLP	G	0.932$\pm$0.001	0.562$\pm$0.010	0.922$\pm$0.001	0.587$\pm$0.014	0.923$\pm$0.000	0.658$\pm$0.006
MLP	SG	0.928$\pm$0.001	0.574$\pm$0.011	0.923$\pm$0.000	0.574$\pm$0.007	0.925$\pm$0.001	0.673$\pm$0.004
MC	SG	0.928$\pm$0.002	0.554$\pm$0.007	0.924$\pm$0.001	0.590$\pm$0.003	0.924$\pm$0.001	0.685$\pm$0.005
CFL	G	0.950$\pm$0.002	0.627$\pm$0.013	0.950$\pm$0.000	0.678$\pm$0.003	0.938$\pm$0.001	0.751$\pm$0.009
HiCFL	G	**0.965**$\pm$0.003	**0.645**$\pm$0.014	**0.957**$\pm$0.001	**0.691**$\pm$0.003	**0.942**$\pm$0.002	**0.769**$\pm$0.006
LR	G	0.860$\pm$0.000	0.206$\pm$0.000	0.788$\pm$0.000	0.280$\pm$0.000	0.833$\pm$0.000	0.429$\pm$0.000
MLP	S	0.804$\pm$0.003	0.144$\pm$0.003	0.767$\pm$0.009	0.260$\pm$0.005	0.804$\pm$0.006	0.421$\pm$0.010
MLP	G	0.836$\pm$0.005	0.234$\pm$0.016	0.813$\pm$0.006	0.295$\pm$0.011	0.842$\pm$0.003	0.467$\pm$0.011
MLP	SG	0.844$\pm$0.003	0.230$\pm$0.015	0.796$\pm$0.008	0.291$\pm$0.011	0.839$\pm$0.006	0.463$\pm$0.013
MC	SG	0.852$\pm$0.006	0.251$\pm$0.019	0.795$\pm$0.014	0.303$\pm$0.017	0.861$\pm$0.004	0.547$\pm$0.006
CFL	G	0.918$\pm$0.001	0.441$\pm$0.009	**0.897**$\pm$0.002	0.408$\pm$0.004	0.887$\pm$0.002	0.563$\pm$0.018
HiCFL	G	**0.940**$\pm$0.008	**0.508**$\pm$0.026	**0.897**$\pm$0.004	**0.421**$\pm$0.014	**0.915**$\pm$0.002	**0.648**$\pm$0.009

w/o PU setting

w/ PU setting

Hierarchical Learning is helpful!

Comparison to Human Performance

	ML-AI	ML-CS	AI-CS
Human	$0.698_{\pm 0.087}$	$0.846_{\pm 0.074}$	$0.716_{\pm 0.115}$
HiCFL	$\mathbf{0.854}_{\pm 0.017}$	$\mathbf{0.932}_{\pm 0.007}$	$\mathbf{0.768}_{\pm 0.023}$

Let humans (5 senior students majoring in CS) and machines judge which term in a query pair is more relevant to ML

HiCFL far outperforms human performance!

Case Studies

The depth of the background color indicates the domain relevance. The darker the color, the higher the domain relevance (annotated by the authors); * indicates the term is a core term, otherwise it is a fringe term.

1-10	101-110	1001-1010	10001-10010	100001-100010
supervised learning*	adversarial machine learning*	regularization strategy	method for detection	tumor region
convolutional neural network*	temporal-difference learning*	weakly-supervised approach	gait parameter	mutual trust
machine learning*	restricted boltzmann machine	learned embedding	stochastic method	inherent problem
deep learning*	backpropagation through time*	node classification problem	recommendation diversity	healthcare system*
semi-supervised learning*	svms	non-convex learning	numerical experiment	two-phase*
q-learning*	word2vec*	sample-efficient learning	second-order method	posetrack
reinforcement learning*	rbms	cnn-rnn model	landmark dataset	half*
unsupervised learning*	hierarchical clustering*	deep bayesian	general object detection	mfcs
recurrent neural network*	stochastic gradient descent*	classification score	cold-start recommendation	borda count*
generative adversarial network*	svm*	classification algorithm*	similarity of image	diverse way

Machine Learning (HiCFL)

Important concepts such as supervised learning, deep learning are ranked very high

Terms ranked before 1010th are all good domain-relevant terms

Case Studies

Given positives (10): deep learning, neural network, deep neural network, deep reinforcement learning, multilayer perceptron, convolutional neural network, recurrent neural network, long short-term memory, backpropagation, activation function.

1-10	101-110	1001-1010	10001-10010	100001-100010
convolutional neural network*	discriminative loss	multi-task deep learning	low light image	law enforcement agency*
recurrent neural network*	dropout regularization	self-supervision	face dataset	case of channel
artificial neural network*	semantic segmentation*	state-of-the-art deep learning algorithm	estimation network	release*
feedforward neural network*	mask-rcnn	generative probabilistic model	method on benchmark datasets	ahonen*
deep learning*	probabilistic neural network*	translation model	distributed constraint	electoral control
neural network*	pretrained network	probabilistic segmentation	gradient information	runge*
generative adversarial network*	discriminator model	handwritten digit classification	model on a variety	many study
multilayer perceptron*	sequence-to-sequence learning	deep learning classification	model constraint	mean value*
long short-term memory*	autoencoders	multi-task reinforcement learning	automatic detection	efficient beam
neural architecture search*	conditional variational autoencoder	skip-gram*	feature redundancy	pvt*

Deep Learning (HiCFL, PU Learning)

Unlabeled positive terms like artificial neural network, generative adversarial network are ranked very high

Terms ranked 101st to 110th are all highly relevant to DL;
Terms ranked 1001st to 1010th are related to ML

Conclusion

- We propose to measure ***fine-grained domain relevance*** — the degree that a term is relevant to a given domain (broad or narrow)
- To handle long-tail terms, we design a novel **core-anchored semantic graph** to bridge domain relevance of terms
- To leverage the graph and domain hierarchy, we propose **hierarchical core-fringe learning**
- To reduce human efforts, we employ **automatic annotation** and **hierarchical positive-unlabeled learning**
- Extensive experiments demonstrate that our methods outperform strong baselines and even surpass professional human performance

Email: jeffhj@illinois.edu
Code and data: https://github.com/jeffhj/domain-relevance

Thanks!

center for
cognitive computing
systems research