CODS-COMAD 2020 Notes

1. Kristian Kersting : Keynote talk --- Deep Machines That Know When They Do Not Know
    a) Molina et al. --- Conditional Sum-Product Networks, Probabilistic Variational Auto Encoders, Conditional SPNs
        i. See papers in AAAI, UAI, AISTATS 2018
    b) Deep Autogenerative Probabilistic Model (Kersting, Molina et al.)
        i. Does not make the assumption on the underlying distribution
    c) The Explorative Automatic Statistician (AAAI 2018)
        i. Not making assumption on likelihood function (Selects the likelihood function based on data)
    d) Human-in-the-learning (Natarajan) --- UT Dallas
    e) Deep Probabilistic Programming language (Kingma, Welley 2013, Razende 2013)
    f) Unsupervised Scene Understanding --- Kersting, ICML 2019
        i. Codebase : github.com/stelzner/supair
    g) Temporal domain : Unsupervised Physics Learning (Kersting, ICLR 2020)
        i. Putting structure and tractable inference into deep models
    h) Co-adaptive ML (Kersting AIES 2019)
        i. Probabilistic (and causal models) are whiteboxes
    i) Computational CogSci (Future directions) --- Josh Tanenbaum
2. Pooled-BiLSTM inspired on ULMFIT
3. EEG paper : Effective Connectivity --- Time domain Granger Causality, Direct Transfer Function, Partial Direct Transfer
    a) Lasso Regression, Sparse Regression, and Yule Walker
    b) SVAR is a promising approach for understanding causal influences
    c) Volume conduction effects
4. Optum - healthcare industry paper (Suman Roy)
    a) Use cases : Prediction of missing attribute classes, Wrong entry detection (attributes mismatch), and Semantic relatedness
    b) Skip-thought vectors, Universal Sentence Encoder
5. Koninika Pal - OpenIE systems
    a) Entities extracted when OpenIE needs to be canonicalized, because of the different surface forms
    b) Context also influences this canonicalization procedure
    c) Baselines : KB-embeddings and Rule Mining
    d) Dataset : DBPedia, Quasimodo
6. Fine-grained Relation Extraction
    a) Canonicalization of relations is necessary\
    b) Hierarchy of relations is missing
    c) Dataset : DBpedia, Infobox, WikiData
    d) Hierarchical Relation Extraction
7. Learning fron Weights: Cost Sensitive Approach for Retrieval
    a) Sponsored Search
    b) Semantic vector represenation or space (only text)
    c) Paper : Shen et al. A Latent Semantic Model with Convolutional-Pooling for IR, 2014
    d) Long tail nature of queries --- Cost-sensitive learning
    e) Metric : Bounce Rate
8. IIT Delhi -- Amitabha Bagchi paper
    a) Kraska et al 2018 --- Learned Bloom Filter
        i. Drawbacks : Inability to adapt to changing distribution

    ii. Inability to adapt to query dynamics
  b) Bloom Filter : No false negatives, Possible false positives
  c) Classifier-adaptive methods vs. Index-adaptive methods
9. Meta-Learning for Few-shot Time Series Classification (TCS Research)
  a) Reference : Finn et al. Model-agnostic meta-learning (JMLR 2017, it is an optimization algorithm, other examples include RepTile), ConvTimeNet
  b) Tang et al., Few-shot Time Series Classification, 2019
  c) Triplet loss (Schroff et al; Facenet)
10. Attributed Multiplex Networks
  a) Set of attributed networks in which each network represents a different type of interactions between the same set of nodes
  b) Normality (Perozzi et al.) = Internal Consistency + External Separability
11. Multi-label Supervised Classifier Learning
  a) Problem transformation : Binary Relevance, Classifier Chains
  b) Algorithm Adaptation Methods
  c) Problems of Label Noise (50% noise or class-conditional noise)
    i. Uncertainty of sources
    ii. Do not have information about : noisy rates and noisy samples
  d) Risk Minimization framework
  e) Learning under Label Noise
    i. Robustness of Risk Minimization
    ii. Defines "Symmetric Label Noise"
  f) Loss functions - Multi-Label
    i. Binary Cross-entry, MAE
12. Actively ranking
  a) Humans are better at ranking pair-wise rather than ranking a set of movies
  b) Rank centrality --- pairwise comparison count matrix
13. Prototype Selection and Dimensionality Reduction on Multi-Label Data
  a) Binary relevance : Treating each class label as a binary classification problem
  b) Prototype selection
    i. Condensation method
    ii. Edition method
14. Vanilla Lift and Shift models usually do not perform well
15. Causal Inference tutorial : WSDM tutorial
https://causalinference.gitlab.io/wsdm-tutorial/intro.html
  a) A toolbox made available by Microsoft named "dowhy"
16. Invited Talk by Amit Sheth : Knowledge Graphs and Big Data
  a) Knowledge-infused Learning
  b) 5% of Google queries is health-based. Google create a separate KG for health queries
  c) Multimodal Knowledge Graphs
  d) Linkedin articles : 15 years of search and knowledge graphs
  e) Knoesis.org/node/06222 : Enriching existing KGs
  f) Ignoring implicit entity extraction is not possible or covered in our recent methods
  g) Understanding and analyzing drug abuses related discussion on web forums ---
    i. Need of Drug Abuse Ontology
    ii. Leads to improvement in recall or coverage
  h) Semantic, Cognitive and Perceptual computing: Advances towards Computing for Human Experience (Amit Sheth)
  i) Gaur Manas et al. "Empathi: an ontology for emergency mapping" : Shallow infusion of using ontology to improving embedding representation

j) Let Me Tell Me About Your Mental Health (CIKM 2018) -- Using Reddit Data, along with medical knowledge bases + Knowledge hierarchy improves performance
k) External Knowledge through Learnable Constraints
l) Mental Health : Bhatt et al. (IEEE 2018 Web Intelligence)
m) AAAI-MAKE 2020 : Using KG for improving embeddings for Autonomous Driving
n) Shallow Infusion, Semi-deep infusion, Deep Infusion
o) AAAI 2020 : Knowledge Infusion : Knowledge-infused Learning Layer

17. Panel discussion : Rapid Proliferation of AI agents "Shourjya Roy, Director of American Express"
    a) In Finance domain, alternate data like income statement, consumer ratings of drivers(indicate professionalism), to estimate the credit-score of the poor population, may lead to financial inclusion
    b) Can be done on a large scale, even with comparable accuracy
18. Probaility mapping functions : Spherical softmax
    a) Need for sparse distribution maps -- sparse attention
    b) Sparsity not in parameter (Weight matrix, which is the general notion). This line of work enforces sparsity in output
    c) Sparsemax (ICML 2016) --- From softmax to sparsemax: a sparse model of attention and multi-label classification
    d) IBM Research contribution : Controllable sparsity --- uses some form of regularization to enforce sparsity
        i. Sparsegen (Control 1) : Increase width ot coverage of non-sparse regions in z1-z2 space
        ii. Sparsity control 2 : Control tge shape of the non-sparse region
    e) Desirable properties of probability mapping functions (See image)
        i. Scale invariance
    f) Sparsity control in multilabel classification
        i. Formulation of multilabel classification is little different
        ii. Aim to get lower number of non-zeroes
19. Zero-shot task transfer (CVPR 2020 Oral) Prof. Vineeth
    a) Assume that all the model parameters lie in a common meta-manifold
    b) Already have a correlation function among the unknown task and known task (pair-wise)
        i. Develop a task correlation matrix, through crowdsourcing (Scale -1 to 3, aggregated using Dawid-Skeene's algorithm)
        ii. Using different methods of deriving task-correlation matrix
    c) Based on the CVPR 2018 (Taskonomy dataset) : Develop a encoder-decoder framework -- Derive task relationship between the encoder of a task and try to predict the decoder of the other task
    d) Task2vec : ECCV 2019 or 2020 --- Came up with a vector representation of the task
    e) Their contribution : Develop TTNet (vineethnb@iith.ac.in)
        i. Along with MSE, we introduce a "data consistency loss"
        ii. See "Implementation" image
        iii. Novel training methodology : at training time use the "self-mode"; at inference time : use transfer mode
        iv. Why better than supervised learning --- this works whenn the tasks are heavily correlated
        v. Deep encoder networks and very shallow decoders
        vi. Codebase : github.com/ArghyaPal/Zero-shot-task-transfer
20. Vidyut Vanika --- Smart Grid --- finished runners-up
    a) AAAI 2019 and AAAI 2020 paper : Reinforcement learning and game theory

b) PowerTAC is a simulation platform that replicates smart grid

21. Deep RL for Syntactic Error Repair (Shirish Shevade)
    a) See Related Works (image)
    b) Codebase available
22. EWISE-Zeroshot learning for WSD (ACL 2019) --- Sawan Kumar (IISc, Bangalore)
    a) Attentive Context Encoder (hypernym, hyponym,), Definition Encoder, Knowledge Graph Embeddings
    b) Approaches trained on training corpus -- cannot handke unseen words, use backoff to address unseen words
    c) Can the system handle rare senses?--- Analysis done
    d) Ablation: Sense embeddings is useful or not --- Ablation analysis is done
    e) Can we learn from limited data : Outperform definition, knowledge and supervised approaches
23. Dr. Geeta Manjunath --- Niramai Healthcare --- Early detection and awareness
    a) 500K deaths all over world, 50% in India
    b) Preventable death
    c) Doctors/radiologists detect through bare eyes --- density difference very less -- attacking younger women --- experience not suitable -- more less when it looks normal
    d) Measuring temperature differences --- no exposure to radiation
    e) ML to reduce false positives
    f) Vascularity analysis --- close to chest walls
24. Panel discussion : ML Research in the Wild
    a) ML Academic Research and System Building --- derive research questions from the deployment and system-building
    b) Algorithms is a small piece of the entire building
    c) Problem that you solve will be there for a few years
    d) Audio and video in local languages --- make the output of the model in a consummable (locally) format
        i. Quantifying whether the content is being consummable --- metrics to see whether a content is read
    e) Whether the AI-driven application are really AI or not. Or is it rule-based --- How to update the model or AI system to update after deployment, after one or two years
    f) Sampling -- how to collect the data that we use for training
    g) Government to collect the data, surface the data available, which is currently stored in silos --- Current efforts to remove barriers to the accessibility of the data
    h) MVPs --- end-user scenario --- people are okay with failures, but not comfortable with undefined outputs
        i. This is a probabilistic model and not deterministic\
    i) Role of ML architects --- to explain the technical details from the data scientist perspective, and explain to the sales and business units
    j) In case of the data-driven product, requires a different framework than the Agile
        i. The two-week deadline, in terms of completion percentage, will not work
        ii. The real challenge comes when the ML products after deployment --- it requires significant time to address these issues
    k) It is important for the version 1 product to be good or usable state, for the Subject Matter Experts (SME) to have an incentive to invest its time and interact with the system, to give a feedback
    l) Human-in-the-loop learning --- iterative process of developing
        i. Also
    m) There will come a stabilization point, where we can get modified

   i. SLA --- some kind of metric for the current siftware product that you are trying to develop
   ii. Calculating the Value At Risk (VAR) for a financial security

25. Tutorial : Repeat of 3 hr-long EMNLP 2019 tutorial (Graphs in NLP) Tutorial link: https://github.com/svjan5/GNNs-for-NLP
   a) Dependency parse -- a type of graph (unlabeled graph) and labeled graph -- Both are at sentence-level
   b) Semantic Role Labeling --- Also at sentence-level
   c) But need not be sentence-level --- Coreference resolution --- at socument-level granularity
   d) At a corpus-level granularity -- existence of KGs
      i. Posing a machine translation problem as a graph transformation problem have been explored before
   e) Not clear how to incorporate graph structures to the recent Deep NLP models
   f) Annervaz et al., NAACL 2018 --- Incorporate KG (from external sources) LSTM is performing better
      i. Dataset : News20 data used for text classification, SNLI dataset
      ii. Augment the LSTM with the external KG --- can do better with less amount of data
   g) Learn better word embeddings by putting constraints like hyponymn, hypernym (Vashisth ACL 2019)
   h) Graph-based Semi-supervised (ACL 2012 tutorial)
   i) Graph NNs vs Graph SSL
      i. GraphNN did not change the representation of the nodes or edges (Representation learning)
      ii. Handle arbitrary relationships (other than similarity)
      iii. Implicit regularization
   j) Key properties of CNN -- translation invariance, local level through convolutions
   k) CNN for Graphs is not generalizable since translation and pooling is not clear for graphs
   l) GCN Formulation : Kipf et al. 2016 --- based on Spectral Graph theory (in EMNLP tutorial)
   m) GCN, just before classifier step, initialization from the Bi-LSTM, as the initial embedding setup
      i. Other way maybe to use GCN initially and then follow-up with stacked BiLSTMs
      ii. Intuition : BiLSTM before helps because it will take care of the local context (, whereas the syntactic parses are able to capture the long-range dependencies ---- subject, verb may be nearer, but object may be further away
      iii. Trained as an end-to-end objective
   n) High-level objective of representing the node, any be similar to “struct2vec”
      i. However, struct2vec may have uniform relationships, but GCNs can handle richer relation types and learns during training
   o) Done over the entire vocabulary space --- There are formulations that consist of a number of small graphs for each sentence, for the entire document. --- Other formulations have a single large graph
   p) Capturing the neighborhood context of the node
      i. Self-attention for GCNs (Velickovic , ICLR 2018)
      ii. Confidence-based GCN (Vashisth, AISTATS 2019)
   q) GCNS may be used for Unsupervised Representation Learning
      i. GraphSAGE (Hamilton, NeurIPS 2017) --- 3 neighborhood aggregators --- LSTMs found to be most effective

   ii. Graph Auto-Encoder (Kipf et al, BDL-NeurIPS 2016) --- VAE-based model --- Objectives are similar (Reconstruction Loss and KL-divergence to stay close to the sinormal distribution
   iii. Deep Graph Infomax (Velickovic, ICLR 2019) --- Readout function contains global information

r) Graph pooling operations is required for representing the entire graph --- for Graph Classification
   i. Graph pooling is NP-hard
   ii. Simple min/max pooling : Inefficient and overlooks node ordering
   iii. Other methods exist, please see EMNLP tutorial

s) GCNs for Directed Labeled Graphs (EMNLP 2017)

t) Hypergraph CNN (NeurIPS 2019)

u) Set of related works, given in the slides

v) Pytorch : DGL

w) Neural Structured Learning in Keras

x) GCNs for Multiple Small Graphs

y) Applications : Words and edge labels are at different semantic spaces

z) Inter-sentence Relation Extraction (Sahu et al. ACL 2019) --- BiLSTM baseline beaten by BiLSTM + GCN

aa) Attention-guided GCN -- Learning to prune us superior to rule-based pruning

ab) Sentence-level syntactic dependencies --- Vashishth et al; SemGCN : Exploits semantics in word embeddings

ac) Summary and Future Directions --- see image

ad) GNN + KG for multi-label image classification

ae) KGs crucial problem is transductive in nature -- for a new entity require a complete re-training
   i. Inductive KG-embedding (Wang et al., AAAI 2019)

af) Open domain KG
   i. Open-Domain QA from KG+Text (Sun et al. EMNLP 2018)

ag) Future work on MUlti-modal KGs

ah) Open Problems and Conclusion
   i. Spectral GCNs rather than first order approximation