Yearly report(2015-01~12)

Tianyi Luo, Tsinghua University






Work done in last year

Research:

- A long oral paper(9 pages) accepted by EMNLP 2015**(CCF B)** as the first author**(5/30/2015)**
- Submit one short paper(4 pages) to EMNLP 2015 as the second author **(6/15/2015)**
- Submit one paper(6 pages) to AAAI 2016 as the equal contribution author **(9/15/2015)**

A long oral paper(9 pages) accepted by EMNLP

- <<Stochastic Top-k ListNet>>
  - **ListNet** is a well-known **listwise learning to rank** model.
  - In this paper, we propose a stochastic **sample** method that significantly reduces the training complexity and better ranking performance.

				P@1			P@10		
Model	Top-k	Sampling	Time (s)	Train	Val.	Test	Train	Val.	Test
C-ListNet	k=1	-	2.509	0.4101	0.4107	0.4119	**0.2684**	**0.2684**	0.2676
S-ListNet	k=1	UDS	0.753	0.4097	**0.4106**	0.4120	0.2680	0.2683	0.2676
S-ListNet	k=1	FDS	0.391	0.4094	0.4090	**0.4127**	0.2679	0.2681	0.2676
S-ListNet	k=1	ADS	**0.375**	**0.4102**	0.4097	0.4121	0.2680	0.2682	**0.2677**
C-ListNet	k=2	-	2275.5	0.4119	0.4043	0.4043	0.2678	0.2674	0.2674
S-ListNet	k=2	UDS	2.898	0.4140	0.4143	0.4130	0.2682	0.2686	0.2681
S-ListNet	k=2	FDS	2.410	0.4145	0.4144	**0.4164**	0.2684	0.2688	0.2684
S-ListNet	k=2	ADS	**2.013**	**0.4162**	**0.4168**	0.4145	**0.2686**	**0.2689**	**0.2687**
S-ListNet	k=3	UDS	4.358	0.4167	0.4204	0.4152	0.2686	0.2681	0.2680
S-ListNet	k=3	FDS	3.997	0.4137	**0.4205**	0.4131	0.2687	0.2695	0.2685
S-ListNet	k=3	ADS	**3.483**	**0.4184**	0.4196	**0.4177**	**0.2692**	**0.2697**	**0.2689**
S-ListNet	k=4	UDS	6.161	0.4145	0.4226	0.4104	0.2686	0.2694	0.2687
S-ListNet	k=4	FDS	5.773	0.4145	0.4232	0.4150	0.2690	0.2695	0.2686
S-ListNet	k=4	ADS	**4.358**	**0.4149**	**0.4247**	**0.4164**	**0.2692**	**0.2700**	**0.2689**

Table 1: Averaged training time (in seconds), P@1 and P@10 on training, validation (Val.) and test data with different Top-k methods. 'C-ListNet' stands for conventional ListNet, 'S-ListNet' stands for stochastic ListNet.

A long oral paper(9 pages) accepted by EMNLP

- <<Stochastic Top-k ListNet>>
  - Significantly reduce the training complexity and get a little better performance.
  - Accepted by EMNLP 2015 (Acceptance rate: 312/1315=24%).

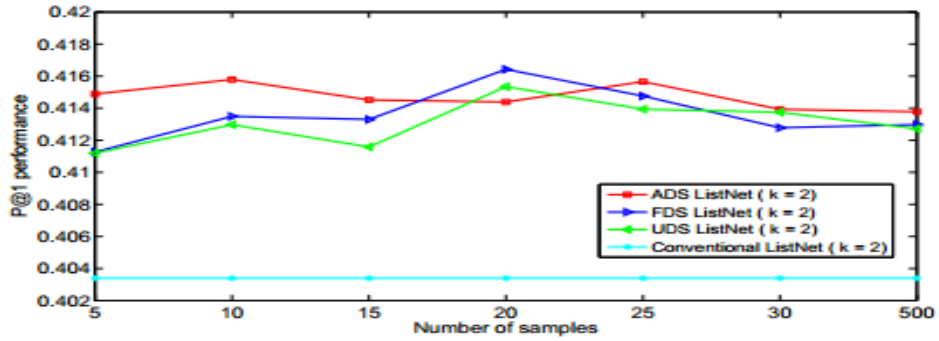
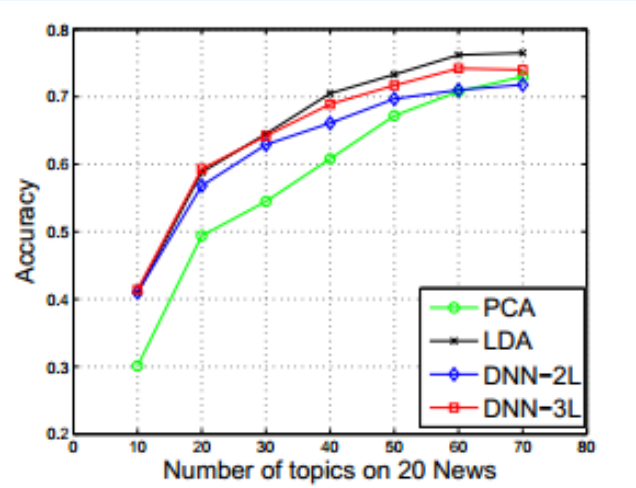


Figure 2: The P@1 performance on the test data with the Top-2 ListNet utilizing the three sampling approaches. The size of the permutation subset varies from 5 to 500.

Submit one short paper(4 pages) to EMNLP 2015

- <<Learning from LDA using Deep Neural Networks>>
  - **Motivated** by the transfer learning approach(**Dark knowledge**) proposed by Hinton et al. (2015), we present a novel method that **uses LDA to supervise the training of a deep neural network (DNN).**



Submit one short paper(4 pages) to EMNLP 2015

- <<Learning from LDA using Deep Neural Networks>>
  - Our experiments on a document classification task show that **a simple DNN can learn the LDA behavior pretty well**, while the inference is **speeded up tens or hundreds of times**.

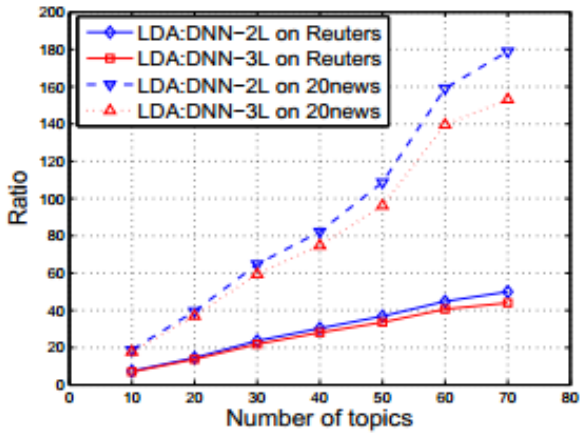


Figure 3: The ratio of inference time of LDA to DNN.

Submit one short paper(4 pages) to EMNLP 2015

- <<Learning from LDA using Deep Neural Networks>>
  - **Topic discovery by transfer learning.** A known advantage of DNNs is that high-level representations can be learned automatically layer by layer. This property may help **DNN to discover topics from the raw TF input.**

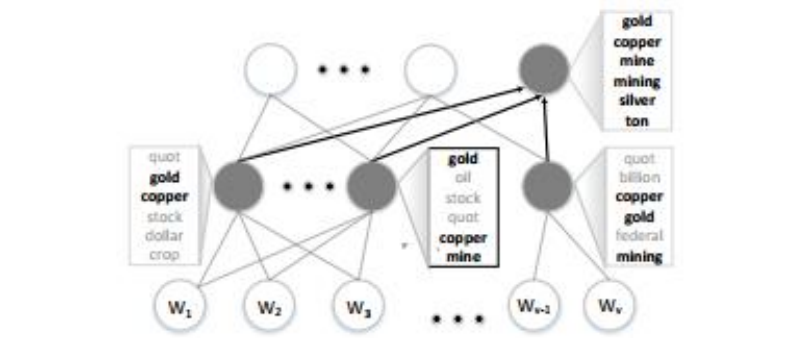
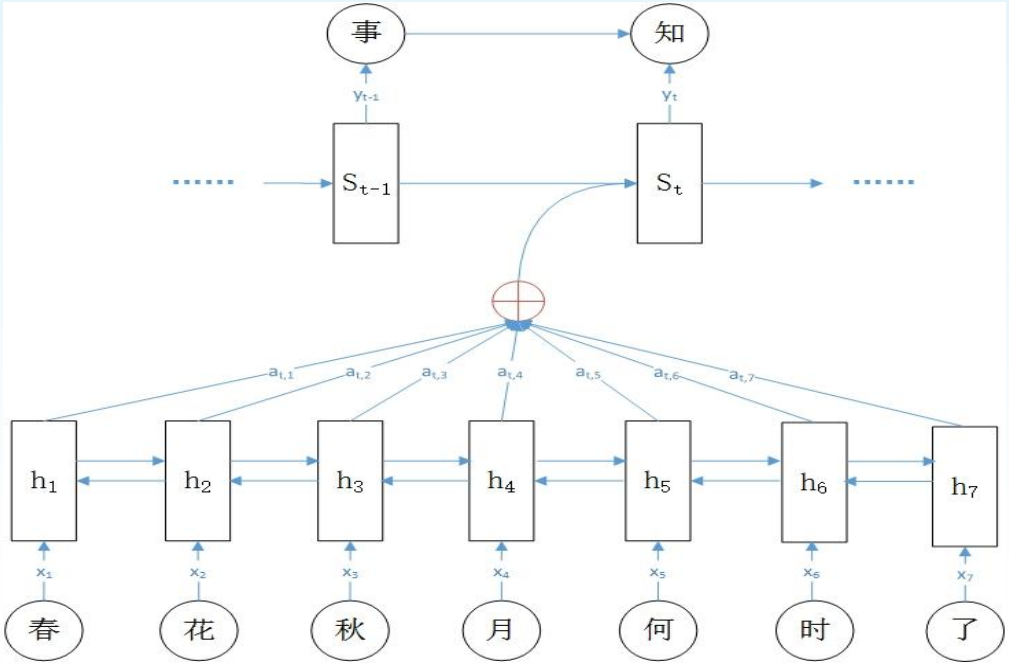


Figure 4: Discovery for the topic 'mining' with DNN. The words in dark are topic related words.

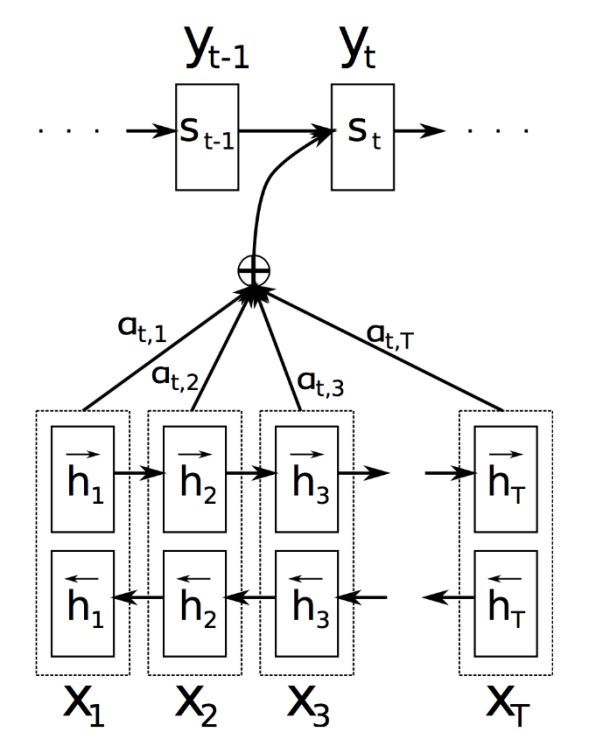
Submit one paper(6 pages) to AAAI 2016

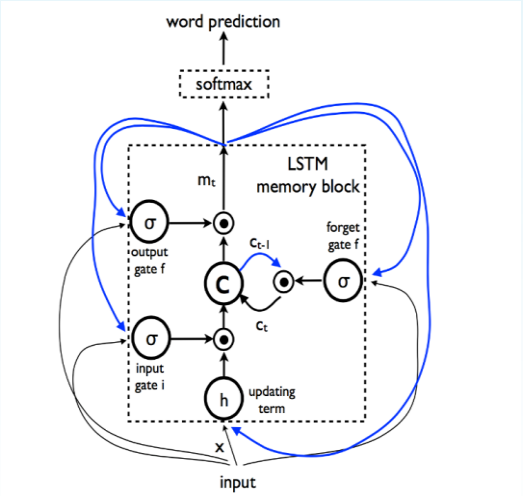
- <<Chinese Song Iambics Generation with Neural Attention-based Model>>



Submit one paper(6 pages) to AAAI 2016

- <<Chinese Song Iambics Generation with Neural Attention-based Model>>





Submit one paper(6 pages) to AAAI 2016

- <<Chinese Song Iambics Generation with Neural Attention-based Model>>

五言唐诗

北国有佳人，
朝中何处深。
泪出三四秋，
东世事多身。

七言唐诗

君问归期未有期，
今夜月江明似衣。
行到城头时看金，
几方飞马雪花枝。

Submit one paper(6 pages) to AAAI 2016

- <<Chinese Song Iambics Generation with Neural Attention-based Model>>

《蝶恋花 * 梦里江河》

万事都归一梦了，
行尽青山，
江上分来尊。
好语燕时难少事，
点破园林无几许。

多闲愁歌中趣了，
两鬓苍苔，
未免教轻寒。
飞入空谷云音路，
不闻何曾相映景。

《虞美人 * 梦里江河》

芙蓉落尽天涵水，
烟水秋平岸。
绿荷多少夕阳中，
背飞双燕贴云独高楼。

一声长合江南枝，
无人早晚来。
流与量酒重携空，
谁家台畔楚宫门阑干。

Work done in last year

Engineering:

- Integrate the learning to rank module to the whole software and get about 4% improvement in the p@1 evaluation.
- Implement Chinese Automatic Error Correction based on language model.
- Implement Chinese Poem, Songci and Couplet generation
- Implement similar questions identification

Implement the learning to rank module

- Offline learning:
  - p@1 准确率： 61%->65%
  - 训练速度：最多提升1000倍
  - 论文发表：EMNLP 2015, 一篇long oral论文
- Online learning:
  - 用户可输入QA对进行系统调教->实时增加QA对
  - 用户对排名靠后结果点赞->实时排名提升

Chinese Automatic Error Correction based on lm

- System design and function:
  - 大语料语言模型（全领域）+小语料语言模型（具体领域）
  - 支持国家领导人名字纠错
  - 支持错别字纠错对手动扩充
- Performance(test set provide by Huilian):
  - 系统标出71处错误（实际错误6处，误报比约9/10）

我进期要申请美国。青华大学。李瑞坏是第十五届中央委员。

一个需要纠错的文本

纠错结果: **1 进期 null 9 青华大学 null 16 坏 环**

Implement Poem, Songci and Couplet generation

- System design and function :
  - 利用attention RNN（循环神经网络）进行唐诗、宋词和对联的自动生成
- Samples:
  - 五言唐诗：北国有佳人（后三句略，请看10页）
  - 七言唐诗：君问归期未有期（后三句略，请看10页）
  - 宋词：虞美人 芙蓉落尽天涵水 （后几句略，请看11页）
  - 对联：生 意 如 春 天，新 行 胜 旧 军
    雄 心 开 伟 业，妙 秋 系 九 州
    文 坛 放 异 彩，艺 生 花 溢 芳

Implement similar questions identification

- System design and function :
  - 利用RNN（循环神经网络）进行相似问句判别
  - 系统用于自动模板扩充和问句相似度判别
  - 训练集：300多万QA对
- Performance:
  - Q:芜湖 的 计算机 软件 水平 考试 在 什么 地方 报名 ？

    A:你 可以 到 安徽 师大呀。

    相似度为：0.823901910035
  - Q:芜湖 的 计算机 软件 水平 考试 在 什么 地方 报名 ？

    A:1 . 找到 卖 你 电脑 发 的 磁盘 2 . 把 磁盘 放 进光 驱 3 . 按照 向导 安装

    相似度为：0.759943815081

Thank You !