# NLPCC-ADL 第二天会议纪要

gaoisbest 自然语言处理实践 2018-08-27

## 1 介绍

Natural Language Processing and Chinese Computing (**NLPCC**) 大会 [1] 正在召开，今天是第二天，主要是 Advanced Disciplines Lectures (**ADL**) 的分享。NLPCC 会持续召开到 8 月 30 号，我也会持续更新每天的报告记录，供大家参考，由于时间紧迫等原因，一些内容只是给出提纲，具体内容以后有时间再详细阐述。

今天有两个主题报告，即神经信息检索 (Neural Information Retrieval, Neural IR) 和 NLP 的联合学习 (Joint Models for NLP)。

## 2 Neural IR

今天上午来自加拿大蒙特利尔大学的聂建云老师 [2] 和其博士生聂一凡共同分享了 Neural IR 的最新研究进展。

### IR 是什么?

给定 Query 和一些 Documents，如何把与 Query 最相关的 Documents 找出来并排序，就是 Information Retrieval 问题。

传统的 IR 方法有 TFIDF, BM25 等，然后利用 Learning to rank (由昨天作报告的刘铁岩老师提出的) 考虑各种特征进行 rerank. 一个缺点是，这些特征一般都是人工指定的，因此出现了 Neural IR，用 Neural Networks 自动提取特征。

IR 与 Sentence matching 是不同的，IR 的每个 document 都是非常长的，而 Sentence matching 都是短文本。

**IR 方法分类**

总的来说，Neural IR 模型可以分为两类，即 **Representation-based** 和 **Interaction-based**. 下面分别介绍两类模型，两类结合的模型，最后总结比较它们。

**Representation-based IR** 也就是基于**表示**的 IR，把 Query 和 Documents 分别表示成 vector，然后做距离度量 (如 cos距离，MLP 等)，常用的模型有 DSSM, CDSSM, ARC-I 等。

**DSSM** 是由微软在 2013 年提出的，本质是 word hashing + DNN, word hashing 用于解决 OOV 问题，并大大降低了维度。Word hashing 是基于 tri-grams 实现的。举例： What a lovely day => [#Wh, Wha, hat, at#, #a#, #lo, lov, ove, vel, ely, ly#, #da, day, ay#]。

DSSM 是用搜索中的用户点击数据训练的，即用户输入 Query，搜索引擎返回一堆 Documents, 如果用户点击了某些 Documents, 这些就是训练时的正样本。值得注意的是，DSSM 的训练数据是 (Query, Document title)，而不是 Document content，因此 DSSM 适用的是短文本匹配。

**CDSSM** 是使用 Convolution 代替 DSSM 的 DNN 来做的，Convolution 的 window size 设置为 3. 效果优于 DSSM.

**ARC-I** 也是使用 Convolution 做 sentence matching 的。

总的来说，对长文本的 Documents 做 representation 是非常困难的，因此 **Representation-based IR 只适用于短文本**。

**Interaction-based IR** 考虑了 Query 的每个词与 Document 的每个词的 Interaction，即如果把 Query 每个词当作列，Document 的每个词当作行，那么二者的 Interaction 会构成一个 matrix。

常用的模型有 MatchPyramid, ARC-II, DRMM 等。

**MatchPyramid** 的原理是使用 Convolution 对 Interaction matrix 做操作。Interaction matrix 的每个元素是 Query 词和 Document 词的 word vector 的 cos 相似度。

**ARC-II** 也是基于 Convolution 的，使用 Conv1D 自动生成 Interaction matrix 的每个元素。

**DRMM** 利用了直方图的概念；**K-NRM** 利用 Kernel pooling；**Conv-KNRM** 基于 Convolution 考虑了 Query 的 n-gram 特征；**SRNN** 加入了空间特征；基于 Document 太长，并不是所有的词语都对 Query 有用的假设，**DeepRank** 只考虑若干个词，即 DeepRank 把 Query 的每个字在 Document 中出现位置的前后几个字拿出来做操作，最后合并这些相关词语。**HiNT** 首先把 Document 均分成固定长度的片段，然后在每个片段是做 SRNN。

**Representation-based IR** 和 **Interaction-based IR** 的结合。

**DuetNet** 的 local model 是考虑 Interaction 的，distributed model 考虑了 representation，然后把两者结合起来。

**二者比较**

由于 IR 问题的 Document 都是长文本，对长文本的 representation 学习一直没有得到很好的解决，因此 **Interaction-based IR** 通常是优于 **Representation-based IR** 的。**Representation-based IR** 的参数很多，而 **Interaction-based IR** 的参数很少。

最后，博士生聂一凡介绍了他的工作，即 Multi-level Abstraction Matching.

实现

Github 的 MatchZoo [3] 提供了上述大部分模型的实现，但是与同学们交流的过程中，说在同样的数据下，这个实现不如自己实现的模型准确度高。🤭

## 2 Joint Models

下午来自浙江西湖高等研究院的张岳老师 [4] 作了关于 Joint Models 的报告。

所谓 **Joint Learning** 就是同时训练多个任务，比如同时训练分词和词性标注。

张老师做的报告非常详细，讲述了从做博士开始发表的关于 Joint Models 的一些 paper 详细原理，由于我是 Joint Models 小白，下午听的也是一脸懵逼 🤦。在这里我就简单列举一下学习资料，希望在大体上有些思路，详细的内容请阅读张老师个人主页列举的 Papers。

**思路**

Joint Learning 的训练思路总体来说可以分为三类：

- Joint Learning, Joint Search
- Separate Learning, Joint Search
- Joint Learning, Separate Search

分类

Joint Learning 的模型设计可以分为两类：

- Statistical Models
- Deep Learning Models

Statistical Methods 分为：

- Graph-Based
- Transition-Based

Graph-Based 分为：

- Joint Label Structure
- Reranking
- Joint Modeling (Multi task)
- Joint Modeling (Single task)

Deep Learning Models 分为：

- Neural Transition-based Models
- Nerual Graph-Based Models (Multi-task Learning)

Nerual Graph-Based Models 分为：

- Cross Task
- Cross Domain
- Cross Lingual
- Cross Standard

中文 NER 的最新研究成果 ACL 2018 paper **LatticeLSTM** [5] 的作者就是张老师。

给我的感觉是张老师对相同研究方向的同事们非常熟悉，无论是澳大利亚，还是印尼等，每个人做的什么工作张口就来，可见张老师在其研究方向上的功力颇为深厚。

## 4 参考文献

[1] http://tcci.ccf.org.cn/conference/2018/index.php
[2] http://rali.iro.umontreal.ca/nie/jian-yun-nie-en/
[3] https://github.com/faneshion/MatchZoo
[4] http://www.wias.org.cn/index.php?a=kydetail&catid=487&id=8925&web=chinese
[5] https://github.com/jiesutd/LatticeLSTM

我们有一个自然语言处理实践交流群，喜欢交流的朋友可以加我微信：gaoisbest 拉大家入群。大家可以关注公众号，阅读 NLPCC 大会的后几天笔记分享。