**2017 知乎看山杯机器学习挑战赛解决方案**

**Team: NLPFakers**

**Rank:4**

# 简介

知乎是世界上最大的中文知识社交平台。知乎自 2011 年创办至今，已经成为一个拥有 2000 万日活用户，每天有数以十万计的高质量的 UGC 内容产生的网站。如何对这些内容进行深层次的理解和高效的分发，是知乎建立的内容生产-内容分发的闭环中的一个重要课题。

目前，知乎上的内容分发的一个重要途径是通过关注关系生成的 Feed 流。关注关系可能是基于人，也可能是基于「话题」标签；从用户关注的话题标签为用户推荐内容，会更加契合用户对不同领域、不同类型的知识的需求。因此，对知乎上的内容进行精准的自动化话题标注，对提升知乎的用户体验和提高内容分发效率有非常重要的支撑作用。因此，知乎算法团队联合中国人工智能学会、 IEEE计算机协会和IEEE中国代表处，举办「知乎看山杯机器学习挑战赛」，期望能够激发更多的优秀人才对于自然语言处理领域的兴趣，并且促进语义分析的应用，从而最终革新甚至改变人的信息获取方式。

参赛者需要根据知乎给出的问题及话题标签的绑定关系的训练数据，训练出对未标注数据自动标注的模型。标注数据中包含 300 万个问题，每个问题有 1 个或多个标签，共计1999 个标签。每个标签对应知乎上的一个「话题」，话题之间存在父子关系，并通过父子关系组织成一张有向无环图（DAG）。由于涉及到用户隐私及数据安全等问题，本次比赛不提供问题、话题描述的原始文本，而是使用字符编号及切词后的词语编号来表示文本信息。比赛还提供字符级别的 embedding 向量和词语级别的 embedding 向量，这些 embedding 向量利用知乎上的海量文本语料，使用 google word2vec 训练得到。除了对原始文本进行大小写转换、全半角转换及去除一些特殊字符（如 emoji 表情、不可见字符）等处理之外，训练数据和预测数据都没有经过任何清洗。

评测方案为预测话题的top5标签的按位置加权的F值。

# Models

**Brief:**

我们所有模型均基于neural network，并且使用了提供的w2v预训练，不同模型以及参数表现如下：

|  |  |  |  |
| --- | --- | --- | --- |
| *single model* | *hidden size* | | LB |
| MLP (word) | | 256 | 0.4031 |
| MLP (word) | | 2000 | 0.4113 |
| MLP (word only title) | | 2000 | 0.4013 |
| TextCNN (word) | | 512 | 0.4102 |
| TextCNN (word + label text) | | 512 | 0.4092 |
| TextCNN (word) | | 2000 | 0.4087 |
| TextCNN (word + char) | | 1024 | 0.415 |
| CNN + att (word + char) | | 1024 | 0.417 |
| RNN (word) | | 1024 | 0.410 |
| RNN + att (char) | | 512 | 0.410~ |
| RNN + att (word) | | 512 | 0.413 |
| RNN + att (word) | | 600 | 0.4142 |
| RNN + att (word) | | 800 | 0.4151 |
| RNN + att (word) | | 1024 | 0.4162 |
| RNN + ATT (DIVIDED) | | 512 | 0.4112 |
| RNN + att (word + char) | | 512 | 0.417 |
| RNN + att (divided) | | 512 | 0.418 |
| RNN + multi att (word) | | 512 | 0.418~ |
| RCNN(char) | | 256 | 0.402~ |
| RCNN(word+char) | | 1024 | **0.422** |
| RCNN + att + hierarchical | | 1024 | 0.418 |
| RCNN + hierarchical | | 1024 | 0.419 |
| RCNN + att (word + char) | | 1024 | 0.418 |
| DMN (word + char) | | 1024 | 0.412 |
| ensemble1 (gzf) | | - | 0.4241 |
| ensemble2 (hhy) | | - | 0.4286 |
| ensemble3 (Jason wu) | | - | **0.4303** |
| Final ensemble | |  | 0.4346 |

**Details**：

1. word char 分别使用了两种不同的最大截断长度word: 158, 256, char: 158, 400.
2. attention layer主要为self-attention, 同时我们又加入了多维度的attention以及description 与 title 之间的attention.
3. 层次化模型以及交互类模型对于性能有较高提升，包括横向（char and word level intersection, hidden intersection）以及纵向（hierarchical concatenation）.
4. 知乎提供了char以及word level的数据供训练，我们通过分别构造char level model, word level model 以及char + word model 来增加模型差异性以及覆盖率.
5. 使用了dropout以及L2 来减少模型的过拟合，BN层未对于我们的模型有显著效果，且收敛速度过慢.
6. 优化选用了Adam(lr=0.00002)以及Adam(lr=0.000085).
7. 实验发现，增大网络的参数能显著提高模型性能，同时需增大正则。
8. finetune对于模型的提升十分显著，我们使用了增大训练数据以及减小lr的方式进行finetune.
9. 使用softmax和sigmoid作为不同的激活函数可以提升模型的diversity，并且在融合的时候对于两类模型分别加权后，normalize sigmoid源的模型，将均值方差于softmax的模型接近再融合可以大大提高融合效果.
10. 由于时间问题，不能够训练足够的char level 模型，Ensemble的时候选择多选一些epoch，并赋予更高的搜索base权重，可以更多的利用模型的差异性提高召回率。

# 总结

感谢知乎、中国人工智能学会、 IEEE计算机协会和IEEE中国代表处共同举办这次比赛，让我们有机会能够接触到真实的数据以及模型的应用场景。在比赛过程中，我们也尝试了各种各样的方法，虽然并不是总能带来提升，但是在对算法的理解上又更近了一步，对自身的能力的提升带来了很大的帮助，期望主办方今后还能举办更多有趣有挑战的竞赛。

# 参考

1. Joulin A, Grave E, Bojanowski P, et al. Bag of tricks for efficient text classification[J]. arXiv preprint arXiv:1607.01759, 2016.
2. Zhang, Ye, and Byron Wallace. "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification." *arXiv preprint arXiv:1510.03820* (2015).
3. Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.
4. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[J]. 2017.
5. Kumar, Ankit, et al. "Ask me anything: Dynamic memory networks for natural language processing." *International Conference on Machine Learning*. 2016.
6. Joulin, Armand, et al. "FastText. zip: Compressing text classification models." *arXiv preprint arXiv:1612.03651* (2016).
7. Lai, Siwei, et al. "Recurrent Convolutional Neural Networks for Text Classification." *AAAI*. Vol. 333. 2015.