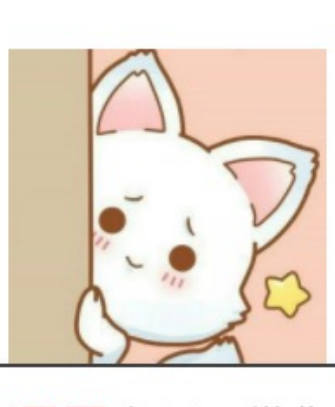
工业界求解NER问题的12条黄金法则

原创 JayLou娄杰 夕小瑶的卖萌屋 2020-07-02 22:20:00 手机阅读



文 | JayLou娄杰，夕小瑶
编 | 可盐可甜兔子酱
美 | Sonata

众所周知，命名实体识别（Named Entity Recognition，NER）是一项基础而又重要的NLP词法分析任务，也往往作为信息抽取、问答系统、机器翻译等方向的或显式或隐式的基础任务。在很多人眼里，NER似乎只是一个书本概念，跟句法分析一样存在感不强。一方面是因为深度学习在NLP领域遍地开花，使得智能问答等曾经复杂的NLP任务，变得可以端到端学习，于是分词、词性分析、NER、句法分析等曾经的显式任务都隐式地编码到了大型神经网络的参数中；另一方面，深度学习流行之后，NER问题相比之前有了比较长足的进步，LSTM+CRF的模式基本成为业内标配，很多人认为“这个事情应该差不多了”。

但！是！

真正在工业界解决NLP业务问题的NLPer，往往发现事情远没法这样轻描淡写。在真实的工业界场景中，通常面临**标注成本昂贵**、**泛化迁移能力不足**、**可解释性不强**、**计算资源受限**等问题，想要将NER完美落（bian）地（xian）可不简单，那些在经典benchmark上自称做到SOTA的方法放在现实场景中往往“也就那样”。以医疗领域为例：

1. 不同医院、不同疾病、不同科室的文本描述形式不一致，而标注成本又很昂贵，一个通用的NER系统往往不具备“想象中”的泛化迁移能力。**当前的NER技术在医疗领域并不适合做成泛化的工具。**
2. 由于医疗领域的严肃性，我们既要知其然、更要知其所以然：NER系统往往不能采用“一竿子插到底”的黑箱算法，处理过程应该随着处理对象的层次和深度而逐步叠加模块，下级模块使用上级结果，方便进行迭代优化、并具备可解释性，这样做可解耦医学事件、便于进行医学实体消歧。
3. 仅仅使用统计模型的NER系统往往不是万能的，医疗领域相关的实体词典和特征挖掘对NER性能也起着关键作用。此外，NER结果往往不能直接使用，还需进行医学术语标准化。
4. 由于医院数据不可出院，需要在院内部署NER系统。而通常医院内部的GPU计算资源又不是很充足（成本问题），我们需要让机器学习模型又轻又快（BERT上不动哇），同时要更充分的利用显存。

以上种种困难，导致了工业界场景求解NER问题时都难以做到BERT finetune一把就能把问题解决，总之

那些口口声声遇事不决上BERT的人们，应该像我一样，看着你们在NER问题上翻车

几天前，卖萌屋的自然语言处理讨论群内就命名实体识别问题进行了一番激烈的讨论，由于讨论持续了接近**2小时**，这里就不贴详细过程了（省略8k字）。

经过一番激烈的辩论，最后卖萌屋的作者杰神（JayLou娄杰）就讨论中出现的若干问题给出了工业界视角下的实战建议（每一条都是实打实的实战经验哇）。

杰神首先分享了他在医疗业务上做NER的七条经验教训：

1. 提升NER性能（performance）的方式往往**不是**直接堆砌一个**BERT**+CRF，这样做不仅性能不一定会好，推断速度也非常堪忧；就算直接使用BERT+CRF进行finetune，BERT和CRF层的学习率也不要设成一样，让CRF层学习率要更大一些（一般是BERT的5～10倍），要让CRF层快速学习。
2. 在NER任务上，也不要试图对BERT进行蒸馏压缩，很可能吃力不讨好。
3. NER任务是一个重底层的任务，上层模型再深、性能提升往往也是有限的（甚至是下降的）；因此，不要盲目搭建很深的网络，也不要痴迷于各种attention了。
4. NER任务不同的解码方式（CRF/指针网络/Biaffine[1]）之间的差异其实也是有限的，不要过分拘泥于解码方式。
5. 通过QA阅读理解的方式进行NER任务，效果也许会提升，但计算复杂度上来了，你需要对同一文本进行多次编码(对同一文本会构造多个question)。
6. 设计NER任务时，尽量不要引入嵌套实体，不好做，这往往是一个长尾问题。
7. 不要直接拿Transformer做NER，这是不合适的，详细可参考TENER[2]。

之后，杰神在群里分享了**工业界中NER问题的正确打开方式：**

非常直接的1层lstm+crf！

> 注：本文所说的lstm都是双向的。

1. ***如何快速有效地提升NER性能？*** 如果这么直接的打开方式导致NER性能达不到业务目标，这一点也不意外，这时候除了badcase分析，不要忘记一个快速提升的重要手段：**规则+领域词典**。在垂直领域，一个不断积累、不断完善的实体词典对NER性能的提升是稳健的，基于规则+词典也可以快速应急处理一些badcase；对于通用领域，可以多种分词工具和多种句法短语工具进行融合来提取候选实体，并结合词典进行NER。此外，怎么更好地将实体词典融入到NER模型中，也是一个值得探索的问题（如嵌入到图神经网络中提取特征[3]）。

2. ***如何在模型层面提升NER性能？*** 如果想在模型层面（仍然是1层lstm+crf）搞点事情，上文讲过NER是一个重底层的任务，我们应该集中精力在embedding层下功夫，**引入丰富的特征**：比如char、bigram、词典特征、词性特征、elmo等等，还有更多业务相关的特征；在垂直领域，如果可以预训练一个领域相关的字向量&语言模型，那是最好不过的了~总之，**底层的特征越丰富、差异化越大越好（构造不同视角下的特征）。**

3. ***如何构建引入词汇信息（词向量）的NER？*** 我们知道中文NER通常是基于字符进行标注的，这是由于基于词汇标注存在分词误差问题。但词汇边界对于实体边界是很有用的，我们该怎么把蕴藏词汇信息的词向量“恰当”地引入到模型中呢？一种行之有效的方法就是**信息无损的、引入词汇信息的NER方法**，我称之为**词汇增强**，可参考《中文NER的正确打开方式：词汇增强方法总结》[4]。ACL2020的Simple-Lexicon[5]和FLAT[6]两篇论文，不仅词汇增强模型十分轻量、而且可以比肩BERT的效果。

   > 将词向量引入到模型中，一种简单粗暴的做法就是将词向量对齐到相应的字符，然后将字词向量进行混合，但这需要对原始文本进行分词（存在误差），性能提升通常是有限的。

4. ***如何解决NER实体span过长的问题？*** 如果NER任务中某一类实体span比较长（比如医疗NER中的手术名称是很长的），直接采取CRF解码可能会导致很多连续的实体span断裂。除了加入规则进行修正外，这时候也可尝试引入**指针网络+CRF**构建**多任务学习**（指针网络会更容易捕捉较长的span，不过指针网络的收敛是较慢的，可以试着调节学习率）。

5. ***如何客观看待BERT在NER中的作用？*** 对于工业场景中的绝大部分NLP问题（特别是垂直领域），都没有必要堆资源。但这绝不代表BERT是“一无是处”的，在不受计算资源限制、通用领域、小样本的场景下，BERT表现会更好。我们要更好地去利用BERT的优势：
   a. 在低耗时场景中，BERT可以作为一个“对标竞品”，我们可以采取**轻量化**的多种策略组合去逼近甚至超越BERT的性能；
   b. 在垂直领域应用BERT时，我们首先确认领域内的语料与BERT原始的预训练语料之间是否存在gap，如果这个gap越大，那么我们就**不要停止预训练**：继续进行领域预训练、任务预训练。
   c. 在小样本条件下，利用BERT可以更好帮助我们解决低资源问题：比如基于BERT等预训练模型的文本增强技术[7]，又比如与主动学习、半监督学习、领域自适应结合（后续详细介绍）。
   d. 在竞赛任务中，可以选取不同的预训练语言模型在底层进行特征拼接。具体地，我们可以将char、bigram和BERT、XLNet等一起拼接喂入1层lstm+crf中。语言模型的差异越大，效果越好。如果需要对语言模型finetune，需要设置不同的学习率。

6. ***如何冷启动NER任务？*** 如果面临的是一个冷启动的NER任务，业务问题定义好后，首先要做的就是维护好一个领域词典，而不是急忙去标数据、跑模型；当基于规则+词典的NER系统不能够满足业务需求时，才需要启动人工标注数据、构造机器学习模型。当然，我们可以采取一些省成本的标注方式，如结合**领域化的预训练语言模型+主动学习**，挖掘那些“不确定性高”、并且“具备代表性”的高价值样本（需要注意的是，由于NER通常转化为一个**序列标注任务**，不同于传统的分类任务，我们需要设计一个专门针对序列标注的主动学习框架）。

7. ***如何有效解决低资源NER问题？*** 如果拿到的NER标注数据还是不够，又不想标注人员介入，这确实是一个比较困难的问题。低资源NLP问题的解决方法通常都针对分类任务，这相对容易一些，如可以采取文本增强、半监督学习等方式，详情可参考《如何解决NLP中的少样本困境》。而这些解决低资源NLP问题的方法，往往在NER中提升并不明显。NER本质是基于token的分类任务，其对噪声极其敏感的。如果盲目应用弱监督方法去解决低资源NER问题，可能会导致全局性的性能下降，甚至还不如直接基于词典的NER。这里给出一些可以尝试的解决思路（也许还会翻车）：
   a. 上文已介绍BERT在低资源条件下能更好地发挥作用：我们可以使用BERT进行**数据蒸馏**（半监督学习+置信度选择），同时利用实体词典辅助标注。
   b. 还可以利用**实体词典+BERT相结合**，进行**半监督自训练**，具体可参考文献[8]。
   c. 工业界毕竟不是搞学术，要想更好地解决低资源NER问题，RD在必要时还是要介入核查的。

8. ***如何缓解NER标注数据的噪声问题？*** 实际工作中，我们常常会遇到NER数据可能存在标注质量问题，也许是标注规范就不合理（一定要提前评估风险，不然就白干了），正常的情况下只是存在一些小规模的噪声。一种简单地有效的方式就是对训练集进行交叉验证，然后人工去清洗这些“脏数据”。当然也可以将noisy label learning应用于NER任务，惩罚那些噪音大的样本loss权重，具体可参考文献[9]。

9. ***如何克服NER中的类别不平衡问题？*** NER任务中，常常会出现某个类别下的实体个数稀少的问题，而常规的解决方法无外乎是重采样、loss惩罚、Dice loss[10]等等。而在医疗NER中，我们常常会发现这类实体本身就是一个长尾实体（填充率低），如果能挖掘相关规则模板、构建词典库也许会比模型更加鲁棒。

10. ***如何对NER任务进行领域迁移？*** 在医疗领域，我们希望NER模型能够在不同医院、不同疾病间进行更好地泛化迁移（**领域自适应**：源域标注数据多，目标域标注数据较少），如可以尝试将特征对抗迁移[11]。在具体实践中，对抗&特征迁移通常还不如直接采取finetune方式（对源域进行预训练，在目标域finetune），特别是在后BERT时代。在医疗领域，泛化迁移问题并不是一个容易解决的问题，试图去将NER做成一个泛化工具往往是困难的。或许我们更应该从业务角度出发去将NER定制化，而不是拘泥于技术导致无法落地。

11. ***如何让NER系统变得“透明”且健壮？*** 一个好的NER系统并不是“一竿子插到底”的黑箱算法。在医疗领域，实体类型众多，我们往往需要构建一套**多层级**、**多粒度**、**多策略**的NER系统。
    a. 多层级的NER系统更加“透明”，可以回溯实体的来源（利于医学实体消歧），方便“可插拔”地迭代优化；同时也不需要构建数目众多的实体类型，让模型“吃不消”。
    b. 多粒度的NER系统可以提高准召。第一步抽取比较粗粒度的实体，通过模型+规则+词典等多策略保证高召回；第二步进行细粒度的实体分类，通过模型+规则保证高准确。

12. ***如何解决低耗时场景下的NER任务？*** 从模型层面来看，1层lstm+CRF已经够快了。从系统层面来看，重点应放在如何在多层级的NER系统中进行显存调度、或者使当前层级的显存占用最大化等。

综上，如果能在1层lstm+CRF的基础上引入更丰富的embedding特征、并进行多策略组合，足以解决垂直领域的NER问题；此外，我们要更好地利用BERT、使其价值最大化；要更加稳妥地解决复杂NER问题（词汇增强、冷启动、低资源、噪声、不平衡、领域迁移、可解释、低耗时）。

除了上面的12条工业界实战经验，群内的一个小伙伴@—— 还提出了一个实际场景经常遇到的问题：

lstm+crf做实体提取时，保证精度的情况下，在提升模型速度上有没有什么好的办法或者建议？

杰神同样给予了一个饱含实战经验的回答：

个人经验来说，1层lstm+CRF够快了。

1. 如果觉得lstm会慢，换成cnn或transformer也许更快一些，不过效果好不好要具体分析；通常来说，lstm对于NER任务的方向性和局部特征捕捉会好于别的编码器。
2. 如果觉得crf的解码速度慢，引入label attention机制把crf拿掉，比如LAN这篇论文[12]；当然可以用指针网络替换crf，不过指针网络收敛慢一些。
3. 如果想进行模型压缩，比如对lstm+crf做量化剪枝也是一个需要权衡的工作，有可能费力不讨好~

可以看出，哪怕是命名实体识别，中文分词甚至文本分类这些看似已经在公开数据集上被求解的任务，放在实际的工业界场景下都可能存在大量的挑战。这也是提醒还未踏入工业界的小伙伴们，不仅要刷paper追前沿，更要记得积极实践，在实际问题中积累NLP炼丹技巧哦。

参考文献

[1] Named Entity Recognition as Dependency Parsing: https://arxiv.org/pdf/2005.07150.pdf

[2] TENER: Adapting Transformer Encoder for Named Entity Recognition: https://arxiv.org/abs/1911.04474

[3] A Neural Multi-digraph Model for Chinese NER with Gazetteers: https://www.aclweb.org/anthology/P19-1141.pdf

[4] 中文NER的正确打开方式：词汇增强方法总结: https://zhuanlan.zhihu.com/p/142615620

[5] Simplify the Usage of Lexicon in Chinese NER: https://arxiv.org/abs/1908.05969v1

[6] FLAT: Chinese NER Using Flat-Lattice Transformer: https://arxiv.org/pdf/2004.11795

[7] Data Augmentation using Pre-trained Transformer Models: https://www.groundai.com/project/data-augmentation-using-pre-trained-transformer-models/1

[8] Better Modeling of Incomplete Annotations for Named Entity Recognition: https://www.aclweb.org/anthology/N19-1079/

[9] CrossWeigh: Training Named Entity Tagger from Imperfect Annotations: https://www.aclweb.org/anthology/D19-1519.pdf

[10] Dice Loss for Data-imbalanced NLP Tasks: https://arxiv.org/pdf/1911.02855.pdf

[11] Label-aware Double Transfer Learning for Cross-Specialty Medical Named Entity Recognition: https://www.aclweb.org/anthology/N18-1001/

[12] Hierarchically-Refined Label Attention Network for Sequence Labeling: https://www.aclweb.org/anthology/D19-1422.pdf