NLP方向总结–翻译

- 论文
  - Training Neural Machine Translation To Apply Terminology Constraints
    - 术语限制的翻译
      - 定义
        - 类似领域适应的翻译，通过使用自定义的专业术语词典，让模型翻译时遇到专业术语，按照词典的专业术语翻译。例如underfit就要翻译成欠拟合而不是失效
      - 其它实现方法：解码时做限制，弱点是增大计算量
        - GBS:Grid Beam Search
        - DBA:dynamic beam allocation
      - 本文方法：术语翻译作为内联标签被插入，即输入加入单词时术语的特征，不在解码阶段限制
        - 数据处理方法有2种
          - Append：train–by–append
            - 图1：源单词alternates后面加上应该翻译正确的术语Stellvertreter
          - Replace：train–by–replace
            - 图1：源单词alternates被从源句子中直接被替换成Stellvertreter
          - 注意tokenize共享了源和目标语句，所以可以直接在源句子上修改
          - 注意额外的标签：0代表源词（默认），1代表源词，2代表目标词。
            - 被嵌入到16维，然后与子词嵌入相拼接
          - 注意：标注了术语的句子大概占10%，其余保持不变，防止暴露偏差
        - 优缺点
          - 优点：推理速度快，缺点：个人认为不能用其它预训练模型，需要从头训练，耗时较长
  - Unsupervised Domain Adaptation for Neural Machine Translation with Domain–Aware Feature Embeddings
    - 翻译无监督领域适应
      - 定义
        - 训练领域适应的翻译模型时，缺乏平行翻译语料对，我们使用领域内的单语数据进行无监督训练，提高模型翻译效果
      - 本文的方法：DAFE(Domain–Aware Feature Embeddings)
        - 通过将表示分解成不同的部分来执行无监督的领域适应
      - 模型:基于Transformer模型修改
        - DAFE结构
          - base网络：θbase表示
            - 学习不同任务和领域的共同特征，翻译的编码器–解码器结构
          - 领域网络：θτ表示
            - 领域特征嵌入学习
          - 任务网络：θγ表示
            - 任务特征嵌入学习
          - 合并嵌入
            - 领域网络的嵌入和任务网络的嵌入通过加法汇入到base网络，注意：每层都会汇入，如图1
        - 训练目标
          - 翻译的目标：给定源语料，源任务，目标领域，拟合翻译的结果
          - 语言模型目标：目标语料句子被随机丢弃和打乱，让模型尝试还原
        - 实验对比：表1
          - 模型1：复制法: 复制的单语数据模型，(Copy)
          - 模型2：逆向翻译法: 逆向翻译的目标–源模型(Back)
          - 模型3：仅仅使用逆向翻译的数据来训练DAFE(Back+DAFE)
          - 模型4：用DAFE训练的目标–源模型生成逆向翻译数据，用它来训练前向模型(Back–DAFE)
          - 模型5：使用Back–DAFE来生成合成的域内数据，并用逆向翻译的数据来训练DAFE模型(Back–DAFE + DAFE), 效果最好
  - Domain Adaptation and Multi–Domain Adaptation for Neural Machine Translation: A Survey
    - 领域适应
      - 定义
        - 不同的领域
          - 出处–文本的来源
          - 主题–文本的主题（专业术语和语义的转移）
          - 体裁–文本的风格，语法，例如鲁迅的的散文风格
        - 领域适应性解释为任何旨在改善现有系统对某一主题或语言体裁的翻译的方案
          - 通用领域的模型翻译如何能快速能在某一个专业领域进行翻译
      - 分类
        - 单领域适应：模型只能在某个领域翻译良好
        - 多领域适应：在多个领域翻译良好
          - 遇到的问题
            - 灾难性遗忘：在领域A上表现出色的模型在领域B上进行了微调，那么它往往会在领域B上提供强大的翻译性能，而在领域A上则会出现极端的性能下降
            - 过拟合或 "暴露偏差"
              - 暴露偏差:模型训练与预测过程之间的不匹配。在训练时每一个词输入都来自真实样本，但是在推断时当前输入用的却是上一个词的输入
              - 过拟合：在领域B上很好，不过领域B变化了一点，效果就差了很多
      - 现代翻译系统模型
        - 表示特征
          - 单词表：使用子词单词表，例如字节对编码(BPE)
            - 优点：单词表示更密集、更有效
          - 内联Tags: 附加特征
            - 自定义一些标签作为输入特征例如自定义术语翻译一个单词，判断说话人的性别
          - 结合上下文
            - 翻译源或目标的上一句和下一句分别是什么，提取出一些特征
        - 主要使用Transformer结构
          - 编码器
          - 解码器
          - 并行
          - self–attention和编码器解码器之间的attention
            - 残差结构：网络变得更深
            - 堆叠结构
            - LN：层归一化，训练更稳定，Robust
        - 损失
          - 交叉熵损失LCE
          - 标签分布平滑：解决暴露偏差
          - 或者目标函数正则化：L2或Dropout
        - 推理
          - 使用beam search
            - 贪心算法的一种，通过对数似然跟踪前N个部分假设。在每个推理步t，对所有可能的beams单token扩展进行重新排序，并选择更新的前N个，直到所有的beams以EOS token终止或超过预先确定的最大长度。
          - 集成模型
            - 多个模型集成，解码速度降低，性能提高
              - 方法1：集成模型权重调整
              - 方法2：一个模型产生翻译，另一个模型对翻译结果纠正，一般是对专业词汇进行修正
          - checkpoint平均
            - 保存训练中的多个checkpoint，并平均参数
        - 数据
          - 加噪声：提高模型鲁棒性
          - 语料不足：单语生成平行语料
            - 逆向翻译：Back Translation
              - 如果那个人不懂英文，那么就用中文翻译成英文，然后用工具把英文翻译回中文，对比原始中文和翻译完成的中文的差别，逆向翻译将比正译更忠实于源文本，正向翻译中的任何错误（如标点符号问题或拼写错误）都会被仔细复制或指出
            - 正向翻译：Forward Translation
              - 正向翻译将自然数据视为源句，并用现有的源到目标NMT模型生成合成目标句。
          - 半合成数据
            - 即单语数据+翻译词，即一些专业术语对应的翻译词
          - 完全合成双语数据
            - 根据词库需要生成自定义的翻译的平行语料库
        - 域模型结构
          - 域标签嵌入：领域标签也可以直接与词嵌入相结合。
          - 特定领域的子网络
            - 特定领域的解码器
            - 领域分类器
          - 适配器层
            - 冻结预训练模型参数，增加新的层用于领域微调，例如图3的解码器最后一部分。
          - 智能寻找需要更新的参数
            - 找到稀疏的或未充分使用的参数子集以实现最小的模型适应，进行域的微调
        - 训练
          - 解决灾难遗忘：尽量减少模型参数的变化
            - 微调更少的步
            - 冻结参数
            - 参数正则化
            - 知识蒸馏
          - 解决暴露偏差：使用最小风险训练目标MRT，使用BLEU评估
            - BLEU:评估翻译输出的事实标准。其核心思想是汇总机器译文和参考译文之间重叠的单词和短语（n–grams）的数量。BLEU度量的范围是0到1，其中1表示与参考译文的输出相同。
          - 提高模型收敛速度和模型质量
            - 课程学习：从容易到难进行学习
        - 低资源数据训练
          - 通过正向或逆向翻译创建合成数据
          - 低资源语言和高资源语言上联合训练以实现 "相似语言正则化"
        - 文档翻译作为一个领域
          - 还可以通过上下文消除歧义
          - 例如翻译的一个文档的风格是应该是保持相同的
        - 神经性解除学习
          - 即有选择的让模型遗忘部分知识
            - 1.防止隐私和关键数据泄露
            - 2. 过时的旧的知识是可以被遗忘的
  - Vocabulary Adaptation for Domain Adaptation in Neural Machine Translation
    - 单词表适应
      - 前提知识
        - 领域适应
          - 我们通常有大量的一般平行语料，例如中文到英文的国际的会议的语料，我们用这个训练翻译模型，但是这个模型对特定领域的翻译并不好，例如用到医学领域，这时，我们需要领域适应
        - 多领域适应
          - 一个翻译模型可能继续要对医学领域翻译好，也需要对生物领域翻译好，也需要对工程领域翻译好
        - 领域适应的方法之微调
          - 通用领域的模型微调到能够对某个领域翻译也很好的情况
        - 无监督领域适应
          - 这个很神奇，有待研究，只需要单语数据，例如某个单个语言，不需要平行语料库
        - 数据增强方法之逆向翻译Back–translation
          - 对翻译模型效果略有提升
      - 定义
        - 领域适应的翻译任务时由一个域（eg:电影）迁移到另一个域（eg:医学）的训练时，通常他们之间的专业术语单词是不同的，直接进行微调训练，可能会导致性能下降，单词表适应是一种解决这种问题的思路。
      - 以前的方法VS本文方法
        - 1. 以前可能会忽略单词表适应，导致目标领域未知词无法处理，或者词义不能正确的转移（可能不同领域的同一个词代表不同的意思）
        - 2. 但通常使用的是sub–tokenization的方法，例如SentencePiece。这时出现OOV的概率较小，但是还是会阻碍单词的理解和词义的转移。
          - eg子词法：例如，"alloy"→"all "+"o "+"y")
        - 3.本文方法能把传统的微调性能提升3.8个BLEU百分点。方法：从目标域的单语数据中导出的一般单词表嵌入投射到源域嵌入空间的嵌入层。（图1）
          - 优点：
            - 1. 与使用单语语料库的逆向翻译Back–translation相比，单词表适应性表现出额外的改进。
            - 2. 单词表的不匹配阻碍了领域的适应，本文的单词表适应方法提高了BLEU性能。
      - 单词表调整的时期
        - 1.预训练模型阶段
          - 缺点：不够灵活，我们可能不能提前知道要是适应到哪个领域，并且过多的不相干的目标域单词表加入会影响到训练
        - 2. 微调模型阶段
          - 直接法：目标域的词嵌入直接替换预训练的NMT模型的编码器嵌入和解码器嵌入
            - 缺点：预训练模型的嵌入空间与目标域词嵌入空间无关，模型效果差
          - 本文的方法: 是把目标域经过CBOW训练后，把单词表和嵌入层都进行替换
      - 本文方法
        - Step1：归纳目标域嵌入
          - 使用CBOW方法从目标域的单语数据中归纳出目标域词嵌入
        - Step2：跨域投影嵌入
          - 目标域词嵌入投影到预训练好的解码器嵌入空间
            - 方法1：线性不相关的变换
            - 方法2：本文所采用的方法，使用locally linear mapping局部线性映射"（LLM），LLM学习一种投影，在映射后保留原始嵌入的局部拓扑结构（位置关系）
        - Step3：微调
          - 用替换后的单词表和嵌入层进行微调
  - A Simple Baseline to Semi–Supervised Domain Adaptation for Machine Translation
    - 半监督领域适应翻译DAMT
      - 分类
        - 模型为中心
          - 源域的翻译任务+目标域的语言模型
        - 数据为中心
          - 伪平行语料库
            - 复制机制
            - back–translation
              - 逆向翻译首先生成从目标语言到源语言的翻译，然后训练翻译模型，将生成的句子映射到目标方。
      - 域适应模型分类
        - 有监督
          - 源域和目标域都有监督的并行数据，尽管源域的数据量比目标域大很多
        - 无监督
          - 没有源域和目标域的平行数据
        - 半监督
          - 只有源域有平行数据而目标域没有
      - 结构
        - Transformer编码器–解码器
        - 字节对编码（BPE）创建共享子词单词表
        - 共享潜表示：编码器可以将任何源语言的输入映射到一个共享的潜表示空间，解码器参数共享
        - 预训练模型初始化参数
      - 目标
        - 损失1：降噪自编码（DAE）语言模型LM
          - 领域单语数据
        - 损失2：迭代逆向翻译
          - NMT1: s2t
            - 源语言到目标语言
              - 大量领域源语言单语数据
          - NMT2: t2s
            - 目标语言到源语言
              - 大量领域目标语言单语数据
        - 损失3：监督下的机器翻译
          - 领域平行语料数据
      - 结论
        - 预训练模型很重要
        - 更多的非平行数据是更好的
      - 实验的三种模型
        - IBT:迭代逆向翻译
          - 优化公式1和2来训练，即损失1和2
        - IBT+SRC
          - 除了优化方程1和2，我们还利用源域的监督数据对公式3进行了优化
        - IBT+BACK
          - 与IBT+SRC的变体类似，不使用源域数据来解公式3，而是使用上述基线BACK提供的伪并行数据
      - 使用的XLM预训练模型
        - 如果我们想用多语言解决方案，可以使用XLM–R预训练语言模型，这个是有预训练模型的
  - A Survey of Deep Learning Techniques for Neural Machine Translation
    - 深度神经网络翻译
      - 发展
        - 规则翻译
        - 统计机器翻译SMT
        - 神经网络翻译NMT
          - 浅层神经网络
          - 使用了NMT的统计机器翻译
          - 纯的神经网络
          - 注意力机制的神经网络
      - NMT的结构
        - RNN
          - 慢，遗忘
        - CNN
          - 长句子的特征只能在高层中发现，表示能力有限，性能低
        - 编码器解码器
          - 编码器
          - 解码器
          - 推理
            - 常用beam seawrch算法
      - 注意力机制
      - OOV问题
        - 提高计算速度，增大单词表
        - 复制机制的翻译方法，解决OOV
        - 子词单元处理OOV
      - SOTA模型
        - 卷积seq2seq，ConvS2S
        - RNMT+模型
        - 基于transformer的模型
      - 面临挑战
        - 长句翻译
        - 对齐机制（注意力）
        - OOV问题
        - Low–Resource语言翻译
        - 文档级翻译，语言翻译等