Quality Estimation的一些应用(wmt12、 wmt13)

- Some interesting uses of sentence-level quality estimation are the following:

  1. Decide whether a given translation is good enough for publishing as is
  2. Inform readers of the target language only whether or not they can rely on a translation
  3. Filter out sentences that are not good enough for post-editing by professional translators
  4. Select the best translation among options from multiple MT and/or translation memory systems

- Some interesting uses of word-level quality estimation are the following:

  1. Highlight words that need editing in post-editing tasks
  2. Inform readers of portions of the sentence that are not reliable
  3. Select the best segments among options from multiple translation systems for MT system combination

Baseline System(Paper : Findings of the WMT 2018 Shared Task on Quality Estimation)

每个任务都有一个特征的基准集(a baseline set of features)并且举办方提供了工具来提取这些features和其它的QE features. 这些features是用来进行model learning.

1. Sentence-level baseline system (Task 1) : **QUEST++**被用来从src和mt files以及平行语料库中提取17个MT system-indepentdent features. 这些特征被用来在SCIKIT-LEARN tookit中使用Radial Basis Function(RBF) kernel训练一个支持向量机回归(SVR)算法。据说这个系统足够用来预测SMT的post-editing effort， 今年也被用作NMT的基准系统
2. Word-level baseline system (Task 2) : **MARMOT**工具被用来提取baseline features。在先前的研究中有28个特征在word-level QE是具有代表性的，除此之外新增加6个特征， 基准系统通过在CRFSuite tool中使用CRF算法把这个任务建模为序列预测问题。该模型使用被动-主动优化算法进行训练。需要注意的是基准系统仅仅被用来在MT output中对单词预测OK/BAD。**没有基准系统被用来预测missing words or erroneous source words.**
3. Phrase-level baseline system : phrase-level features同样使用**MARMOT**提取。但是与word-level features不同， 它们基于sentence-level features(使用QUEST++工具进行提取)。这些特征并**没有使用机器翻译系统内部的信息(called black-box features)**。另外phrase-level QE被当作一个序列标注任务。建模与训练过程与word-level QE相同。同样基准系统仅仅被用来在MT output中对短语预测OK/BAD

数据集处理(Paper : Findings of the WMT 2018 Shared Task on Quality Estimation)

- Task1 and 2 : 过滤掉大多数hter=0的数据
- Task3 : 该任务使用Task1的一个子集， 里面的所有phrases都被标注为4个标签，这个子集是如何获得的呢？首先Task1的数据已经post-edited了, 从中过滤掉hter=0和hter=0.3及以上的translations. 之后在人工标注预算以内随机选择一个较大的子集。对于phrase labelling, 使用BRAT工具。给专业译者提供要标注的翻译以及相应的源句。标注过程在在线环境下进行，所有的翻译已经做了预处理(在phrase-level上都标注为OK)，译者的任务是修改不正确的短语的label。采用的是"悲观模式"，也就是短语内所有词的label与短语的label(OK、BAD、BAD_ORDER)是一致的。Task3总共有两个子任务，Task3a : **将短语注释的结果扩展到它的所有词上**，该任务被定义为单词级别的预测任务。Task3b : 短语级别的预测。(**关于task3a与task2的区别，第一点是task2提供的label是通过ter工具自动计算得到的，而task3a提供的label从这里看是通过人工完成的。再一点就是task2中没有对词序错误进行标注，而task3a中专门对词序错误进行了标注**)

Sentence-level QE(原始笔记...)

1. Participating systems are required to score (and rank) sentences according to post-editing effort.

2. four language pairs =====> (pbmt or nmt)
3. pre-treated : kept a small proportion of HTER=0 sentences in training, development and test sets.
4. training and development dataset : src / mt / ref / pe / hter / add_labels(including post-editing time in seconds || number of keys pressed for 10 types of keys || post-editor id) =====>(pbmt or nmt)
5. test dataset : src / mt =====>(pbmt or nmt)
6. evaluation :
    - Scoring: Pearson's correlation (primary), Mean Average Error (MAE) and Root Mean Squared Error (RMSE).
    - Ranking: Spearman's rank correlation (primary) and DeltaAvg.

Word-level QE

错误种类有: 替换(substitution)、插入(insertion)、删除(deletion)。(具体看下面的2、3小点)

1. 目标
    - 对翻译的结果在单词级别上进行二元分类(OK or BAD)
    - 在翻译的某个位置上预测是否有单词丢失
    - 在原句中预测可能导致翻译出错的单词
2. 与前几年相同，单词级别二元分类的label(OK or BAD)是通过词对齐自动获得的，词对齐过程是通过TER工具(默认设置且shifts被禁用"-d 0")在机器翻译文件与post-edited文件之间进行的。
    - **Shifts (word order errors) were not annotated as such (but rather as deletions + insertions) to avoid introducing noise in the annotation.**(TER的README文件中关于-d选项的描述为：maximum shift distance, default is 50 words，从"-d 0"这个选项上来看，这里没有进行词序错误的修正)。
    - Target tokens originating from insertion or substitution errors were labeled as BAD. All other tokens were labeled as OK. (在机器翻译语句中，来自于**insertion**或者**substitution**错误的token被标记为BAD，其它的token被标记为OK)。
3. 为了标注**deletion**错误，在每个单词之后以及句子开头添加gap tag, 如果在gap tag这个位置上存在**deletion**错误，则标记为BAD，否则标记为OK。
4. 为了在source sentence中标注与机器翻译中**insertion**与**substitution**错误相关的单词。需要把source sentence与post-edited sentence进行对齐操作，对于每一个在post-edited sentence中的token，如果它在机器翻译语句中被**删除**或者**替换**了，那么相应的已经对齐好的在source sentence中的词被标记为BAD，否则标记为OK
5. 数据集
    - 训练集与验证集 :
        - source sentence
        - machine translated sentence
        - post-edited machine translation
        - reference translation(这个文件感觉上没有作用)
        - src_tags(对应task2中目标的第三个)
        - tags(对应task2中目标的第一和第二个，单词级别的quality labels：首先进行单词级别的二元分类任务，一共有N个token，之后在每个单词之后增加gap token，在句子开头也有一个，共N+1个，gap token相应的标记为OK or BAD)。
        - source-sentence与machine-translation的词对齐文件(包括下面的task3也提供了这个文件，但是整个任务流程中好像并没有用到，需要用到的文件应该是source-sentence与post-edited-sentence的对齐文件)
    - 测试集

        - source sentence
        - machine translated sentence
        - source-sentence与machine-translation的词对齐文件。

6. 结果评估 :

    - 通过OK以及BAD标签上的F1-scores相乘所得, 对于task2来说，三个过程获得的结果的评分过程是相互独立，分别是

        - words in the MT, ('OK' for correct words, 'BAD' for incorrect words)
        - gaps in the MT ('OK' for genuine gaps, 'BAD' for gaps indicating missing words)
        - source words ('BAD' for words that lead to errors in the MT, 'OK' for other words)

    - 2015年之前，一直使用BAD标签上的F1-scores得分，但是这个指标会让QE系统把更多的词标记为BAD，F1-OK以及F1-BAD的相乘结果更具有公平性。(WMT16)

Word/phrase-level QE with human annotation for phrases

1. 目标 : 研究人工标注短语的效果
2. 过程：见数据预处理中的task3

3. task3a与task3b中共用的文件(有一点奇怪的地方是README文件中并没有把source sentence加入共用文件中，但实际上两个文件完全相同，这里我把它归结到共用文件中)

    - 训练集与验证集

        - source sentence(已经分词)
        - source sentence与machine translation的词对齐文件(task3a与task3b的翻译存在着非常细微的区别，所以这里的对齐文件)
        - src_tags文件(对应task2中目标的第三个)

    - 测试集

        - source sentence(已经分词)
        - source sentence与machine translation的词对齐文件

4. task3a

    - 训练集以及验证集

        - machine translation(已经分词)
        - "mt_"tags文件(对应task2中目标的第一和第二个，单词级别的quality labels：首先每个单词根据它所属的短语进行标注，一共有N个token，之后在每一个单词后增加gap token，在句子开头也有一个，共N+1个，如果这个gap token在短语中，那么跟随phrase label标注为OK、BAD、BAD_word_order， 如果在短语之间，则标记为OK或者BAD_omission， BAD_omission指的是那里应该有一个或者多个tokens)

    - 测试集:

        - machine translation(已经分词)

5. task3b :

    - 训练集以及验证集

        - machine translation(以"||"划分成为短语, **与task3a不同**)

        - tags文件(短语级别的quality labels：每个短语被标注为OK、BAD、BAD_word_order，短语之间以及句首的gap tag被标记为OK或BAD_omission)
        - src_phrases(把source sentence文件用||符号划分短语所得到的文件)
    - 测试集
        - machine translation(以”||”划分成为短语, **与task3a不同**)
        - src_phrases(把source sentence文件用||符号划分短语所得到的文件)
6. 结果评估
    - 好像是与task2相同，虽然多了BAD_word_order、BAD_omission, 但是最终对该任务进行评估计算F1-scores的时候好像并没有用到
    - 具体表现在task3a与task2的提交格式完全相同，task3b也仅仅多了BAD_word_order标签。
    - 查看数据会发现task3a与task3b的tags文件中含有这4种标签, 关于task3a中没有BAD_word_order标签的提交，可以理解为单词级别上的错误性预测不需要关心词序错误的问题，

Document-level QE

1. 文档级别的QE，今年新增加的内容
2. 数据集来源于Amzon Product Reviews dataset.
    - 训练集与验证集 (下面的数据为完整的一份数据，共n份)
        - annotations.tsv(与MQM计算相关的信息)
        - source.segments(第一行为商品标题，随后的几行为商品的描述信息)
        - mt.segments(机器翻译结果)
        - total_words(单词的数量，被用来计算MQM)
        - document_mqm(MQM score, 这个得分是通过word-level errors与它们的severity计算而来的)
    - 测试集(下面的数据为完整的一份数据，共n份)
        - source.segments(第一行为商品标题，随后的几行为商品的描述信息)
        - mt.segments(机器翻译结果)
        - total_words(单词的数量，被用来计算MQM)
3. MQM计算
    - An error-free translation scores 100%. To calculate a sentence’s MQM score with standard LISA severity weights the following formula applies: MQM Score (%) = 100 - ((**IssuesMinor** + 5 * **IssuesMajor**+ 10 * **IssuesCritical**)/**Sentence length**)*100. A 28-word sentence with 2 minor issues and 1 major issue would have therefore a score of 100-(2+5)/28*100= 75%. (这里应该没有什么问题, 需要的数据为Minor、Major、Critical、Sentence length，在训练集中，这些数据以及MQM得分都提供了，在测试集中仅仅提供Sentence length)
4. 结果评估
    - Pearson’s correlation between the true and predicted document-level scores.

关于task2、3的数据集

- 需要说明的点：第一个点是task3a的source sentence与task3b的source sentence相同，但翻译却有一点点差异；第二点是task3b的source sentence与source sentence phrase(去掉"||"后)有不小的差异；第三个点是task2与task3都提供了source sentence与machine translated sentence的对齐文件；

- 好像是一旦把句子进行短语级别的划分就一定会多或者少一些单词，应该是划分的过程中自动的修改了源句，那么第一点与第二点就能够解释了；关于第三点对齐文件的作用还没有想明白

问题

- Baseline System 3. 中所谓的”这些特征并没有使用机器翻译系统内部的信息”该如何理解？
- Word/phrase-level QE with human annotation for phrases 6. 结果评估部分对于F1-scores的计算是否只是针对OK与BAD标签
- 其它