医疗机器翻译初步调研结果

-2018.02.26

一、问题描述

提供医疗相关翻译训练数据，自动的给出测试集上的翻译结果。

二、数据集

2.1 WMT16

WMT16[1] 新加入生物医疗翻译任务，旨在翻译生物和健康领域的科学文献，其使用数据集检索自 Scielo[2]。Scielo 是一个生物文献数据库，用以提供发展中国家一种提高知名度和获取科学文献的有效途径。

其包含的语言对有：

1) 英语-法语和法语-英语；
2) 英语-西班牙语和西班牙语-英语；
3) 英语-葡萄牙语和葡萄牙语-英语；

语言对的文档内容为文献的题目、摘要或者是其两者的组合。

2.2 WMT17

WMT17[3] 在之前的生物医疗翻译任务基础上新加了两个数据库的文档，有 1）UFAL[4] 包含不同来源的医疗文本 2）Khresmoi[5] 包含 1500 个医疗相关文献的总结，语言有捷克语，英语，法语，德语，匈牙利语，波兰语，西班牙语，其 BLEU 的评测结果为 [6]。

其提供的语言对有：

1) 英语-捷克(新)；
2) 英语-法语和法语-英语；
3) 英语-德语(新)；
4) 英语-匈牙利语(新)；
5) 英语-波兰语(新)；
6) 英语-葡萄牙语和葡萄牙语-英语；
7) 英语-罗马尼亚语(新)；
8) 英语-西班牙语和西班牙语-英语；
9) 英语-瑞典语(新)；

一些数据集概要为如图所示：

	Czech–English			German–English			French–English		
source	pairs	src	tgt	pairs	src	tgt	pairs	src	tgt
UMLS	70	218	224	86	303	317	80	301	256
DBpedia	69	141	151	306	685	718	375	895	893
EMEA	319	5,400	5,598	347	5,567	5,947	354	7,202	6,068
MuchMore	–	–	–	2	141	148	–	–	–
PatTR	–	–	–	1,594	55,070	58,458	–	–	–
COPPA	–	–	–	–	–	–	1,190	33,729	27,149
Com-Crawl	161	3,542	3,976	2,395	55,989	59,782	3,236	94,040	82,170
EuroParl	627	14,815	17,387	1,866	50,372	52,987	1,958	64,258	55,502
JRC-Acquis	593	18,030	20,737	773	24,347	26,233	781	29,762	25,979
News-Com	140	3,219	3,580	177	4,654	4,635	157	5,080	4,151
OJEU	1,859	44,573	50,176	1,715	41,933	44,851	2,031	64,589	54,776
DBpedia	148	333	360	681	1,562	1,712	745	1,979	1,942
CzEng	10,282	147,549	169,669	–	–	–	–	–	–
PatTR	–	–	–	7,979	290,184	321,412	–	–	–
Linguee	–	–	–	52	70	92	–	–	–
Hansard	–	–	–	–	–	–	837	21,622	18,042
MultiUN	–	–	–	–	–	–	10,267	375,337	310,649
COPPA	–	–	–	–	–	–	7,320	205,735	166,142

三、 评价标准

3.1 人工评价

人工评价[13]有几个权衡因素，准确率和流畅度，机器翻译结果与经验人士的翻译结果越相近越好。

3.2 BLEU

由于人工评估代价太高且无法复用 ，BLEU(BiLingual Evaluation Understudy)[11]用于分析候选译文和参考译文中 n 元组共同出现的程度，量化机器翻译与其最接近的一个或多个参考人工翻译结果的相似度。这种评价方法是独立于语言的，并且和人工评估结果高度正相关，计算代价也很小，然而它并没有考虑到翻译的可理解性与语法相关性。

BLEU 的取值范围为[0, 1]。除非翻译结果与某个参考结果完全一致，否则很难达到 1。此外，往往人工翻译结果也不一定是 1。

每个源语句对应的参考结果越多，BLEU 的得分也会越高，因此在比较不同翻译结果的好坏时，要确保在相同的参考翻译个数的预料集合上。

3.3 NIST

NIST 在 BLEU 的基础上进行了一些修正，会计算一个特殊的 n-gram 包含所有信息量。如果找到了一个正确的 n-gram, 这个 n-gram 越稀少，它的权重也就越大。

3.4 METEOR

METEOR(Metric for Evaluation of Translation with Explicit ORdering) [12] 基于一元模型的精确率和召回率的调和平均数，可以产生在句子或分割层面上与人类评价标准更具相关性的结果，而 BELU 是在整个语料库上面寻求相关性。

其计算公式为：

$$p=\gamma(\frac{c}{u_m})^{\theta}$$

$$P=\frac{m}{w_t}$$

$$R=\frac{m}{w_r}$$

$$F_{mean}=\frac{PR}{\alpha P+(1-\alpha)R}$$

$$M=F_{mean}(1-p)$$

其中，p 为惩罚因子，P 为一元模型精准率，R 为一元模型召回率，M 为最终计算的评价值。常量参数 γ =0.5，θ =3，α =0.9；m 为同时出现在候选翻译和参考翻译中的一元模型的数量，w_t 为候选翻译中的一元模型数量，w_r 为参考翻译中的一元模型数量。

四、方法

4.1 爱丁堡 WMT16

方法 [9] 包含预处理，文字对齐，语言模型，基准特征，微调，解码六个步骤。

1）预处理使用 Moses 工具包来进行归一化标点符号和进行词语切分；

2）文字对齐使用 fast_align；

3）在每一个单一语料库上训练单个模型，然后再对这些模型进行线性插值来获取最终的单一模型；

4）基准特征使用特征的线性加权组合来估计分数翻译假设。这些特征包含 5-gram 语言模型分数，短语翻译和词汇翻译分数，词汇和短语惩罚，一个线性失真分数；

5）由于特征集合太大（一般有 500 到 1000 个特征），使用 k-best batch MIRA 来进行微调；

6）使用 cube pruning 来进行解码。

4.2 爱丁堡 WMT17

方法[10]包含预处理，基础模型和集成方法：

1）预处理使用 Moses 工具包来进行归一化标点符号和进行词语切分；

2）训练一个 attention encoder-decoder 模型来作为翻译模型。具体的模型细节为 word embedding 尺寸为 512，隐藏层尺寸为 1024，优化器为 Adam, 初始学习率为 0.0001, batch size 为 80，每 10000 次迭代就进行一次验证，如果连续 10 次验证集指标没有上升就停止训练。

3）采用两种集成方法。一种是使用最后 N 次 checkpoint 的模型来进行集成，另一种是使用不同的超参然后训练不同的模型来进行集成。

五、现有方法存在的问题

1）可能由于源域语言中的部分词汇或概念无法翻译到目标语言中去，翻译结果中有很多遗漏的词汇或者混合有源域语言中的部分词汇；

2）不正确的形容词和名词顺序；

3）在涉及到性别和数量时，名词、动词和形容词翻译词性不一致；

4）标点符号错误；

5）单词大小写不一致；

6）遗漏首字母缩写词汇的翻译；

五、总结

在医疗行业机器翻译问题上，深度学习时代之前是使用统计机器学习方法来进行估计，最近从论文中看到也是过渡到了神经网络模型。与常规的机器翻译任务相比，医疗行业的机器翻译问题差异体现在数据集上，其数据集主要为医疗行业文献题目、摘要、或者其两者的混合，存在着一定的局限性。

参考资料

1. http://www.statmt.org/wmt16/biomedical-translation-task.html
2. http://www.scielo.org/php/index.php
3. http://www.statmt.org/wmt17/biomedical-translation-task.html
4. https://ufal.mff.cuni.cz/ufal_medical_corpus
5. https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2122
6. http://www.statmt.org/wmt17/wmt-2017-biomedical.pdf
7. https://cris.fbk.eu/retrieve/handle/11582/307240/14326/W16-2301.pdf
8. http://www.aclweb.org/anthology/W17-4719
9. http://www.aclweb.org/anthology/W16-2327
10. https://arxiv.org/pdf/1708.00726.pdf
11. Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting o

n association for computational linguistics. Association for Computational Linguistics, 2002: 311-318.
12. Denkowski M, Lavie A. Meteor universal: Language specific translation ev aluation for any target language[C]//Proceedings of the ninth workshop on statistical machine translation. 2014: 376-380.
13. https://en.wikipedia.org/wiki/Evaluation_of_machine_translation