顶会期刊

- TOSEM(Transactions...)
- TSE(Transactions...)
- ASE
- ICSE
- ISSTA
- FSE/ESEC

Metamorphic Robustness Testing of Google Translate

- 论文pdf
- 摘要：当前有关机器翻译软件测试的研究主要集中在有效，格式正确的输入的功能正确性上。相比之下，健壮性测试通常被忽略，后者涉及软件处理错误或意外输入的能力。 在本文中，我们建议解决这一重要缺陷。 使用变体鲁棒性测试方法，我们将原始输入的翻译与后续输入的翻译进行了比较，这些输入具有不同类别的次要错字。 我们的经验结果表明，Google Translate缺乏鲁棒性，从而为神经机器翻译的质量保证开辟了新的研究方向。
- 关键字：鲁棒性测试,准则问题,蜕变测试,蜕变鲁棒性测试,机器翻译,MT4MT

  ps:MT4MT 蜕变测试for机器翻译,这个关键词实在是太符合了,且关于该方向的论文文献貌似有很多.这之前搜索到的文献整理完毕之后将会着重搜索该方面的文献

- pub info:Dickson T. S. Lee, Z. Q. Zhou, and T. H. Tse. 2020. Metamorphic Robustness Testing of Google Translate. IEEE/ACM 5th International Workshop on Metamorphic Testing (MET'20). In Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops (ICSEW'20), ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3387940.3391484

- 文章侧重点: 该篇论文主要侧重于常见输入错误,例如轻微的错别字\大小写错误\句末标点符号的省略\标点符号轻微错误\句子中包含的轻微噪音\更改句子中的主语或者宾语\同义动词替换等,对于机器翻译软件-针对于GOOGLE TRANSLATE--鲁棒性的影响,并且对每一个输入错误类型提出一个鲁棒性相关的蜕变关系(MR),并且支出谷歌翻译的常见翻译错误类别.
- 参考价值:提出的6个MR准则,没有技术上有价值的参考作用,这居然也能发论文...

A Monte Carlo Method for Metamorphic Testing of Machine Translation Services

- 论文PDF
- 摘要:随着机器翻译服务的日益普及，能够评估其质量变得越来越重要。但是，测试准则问题使自动测试变得困难。在本文中，我们提出了一种结合蜕变测试的蒙特卡洛方法，以克服测试准则问题。使用这种方法，我们评估了三种流行的机器翻译服务的质量-Google Translate，Microsoft Translator和Youdao Translate。我们将源语言设置为英语，目标语言包括中国语，法语，日语，韩语，葡萄牙语，俄语，西班牙语和瑞典语。收集了33,600个观测值的样本（总共涉及100,800个实际翻译），并对翻译进行了3×56析因设计分析。基于这些数据，我们的模型针对所考虑的每种目标语言发现Google表现最好(就所使用的蜕变关系而言)。还确定了印欧语系语言有产生更好结果的趋势。
- 关键词:Machine translation quality, oracle problem, metamorphic testing, Monte Carlo method, natural languages ->机器翻译质量,准则问题,蜕变测试,蒙特卡洛方法,自然语言
  ps:蒙特卡洛方法->统计模拟方法,以概率统计理论为指导的数值计算方法,是指使用随机数（或更常见的伪随机数）来解决很多计算问题的方法。
- pub:2018 ACM/IEEE International Workshop on Metamorphic Testing
- 文章侧重点:提出了两个研究问题:
  是否可以在缺少测试准则的情况下进行机器翻译服务的测试;
  若上一问的回答是肯定的,那么在变质测试框架的背景下，我们的方法在何种程度上可以区分好的翻译服务和差的翻译服务？
- 按照解决以上两个问题来架构后续文章,首先提出本文基于的MR关系和测试方法. 首先提出一个以往的做法,MR是RTT,指将原始句子S进行一个双向翻译之后得到的结果S',与S进行相似度上的对比,该方法有明显的缺陷,没有测试同一个系统而是测试两个系统;于是提出了一个基于单向翻译的non-RTT方法的MR,即完美的翻译器直接(原始语言->目标语言)翻译结果或者是间接(原始语言->中间语言->目标语言)翻译结果应当是相同的,
- 文中给出的公式就是$P_L = P'_L$,P_L表示直接翻译结果,P'_l表示间接翻译结果,这之后会使用一个蒙特卡洛方法来进行相似性(一致性,软件质量的表现)的对比评估.评估指标如下表,每个指标给出0/1的结果,1表示完美配对,0表示否然,取三个指标的平均数作为每个对比的记录结果(得分)

指标名称	相关描述简介
Levenshtein Distance	编辑距离
BLEU	nltk库提供
Cosine Similarity	两个向量化的句子的余弦距离

- 实验设置中包含了本文其他侧重点,是针对间接翻译引出的变量path,由不同间接翻译的方式决定.最后每个样本的具体得分则由翻译器种类和间接翻译的path,二者交互项的因素以及修正误差等来综合判定.文末有一个模型的诊断,以及模型意义的分析,该部分有创新性的参考价值。

Metamorphic Testing for Machine Translations: MT4MT

- 论文pdf
- 摘要:自动化机器翻译软件和服务已变得广泛可用并日益流行。由于自然语言的复杂性和灵活性，这种软件的自动测试和质量评估非常具有挑战性，尤其是在没有人工准则或参考翻译的情况下。此外，即使可以使用参考翻译，某些主要的评估指标（例如BLEU）对于短句也不可靠，短句现在是Internet上流行的句子类型。为了缓解这些问题，我们一直在使用蜕变测试技术以全自动的方式测试机器翻译服务，而无需任何人工评估人员或参考翻译的参与。本文报告了我们的进展，并提出了一些有趣的初步实验结果，这些结果揭示了两种主流机器翻译服务（谷歌翻译和微软翻译）中英译汉的质量问题。这些初步结果证明了蜕变测试对于自然语言处理领域中的应用程序的有用性和潜力。
- pub:2018 25th Australasian Software Engineering Conference (ASWEC)
- 关键词:Machine translation, software testing, quality evaluation, oracle problem, metamorphic testing, MT4MT
- 侧重点:这篇文章基于上一篇论文(蒙特卡洛方法)做出延伸讨论,研究问题是:是否可以开发新的MT4MT技术来自动检测主要机器翻译服务中的实际缺陷？文章主体部分有七个模块,其中三个为主要的模块.分别是背景和关键概念的理解,解释本文采取的测试方法主要内容,实验设计分析实验结果,对比本文的研究与相关工作,总结本文论点以及未来发展方向.
  背景介绍：测试方法(与上文相同)–一个完美的翻译器直接(原始语言->目标语言)翻译结果或者是间接(原始语言->中间语言->目标语言)翻译结果应当是相同的.这个验证方法的条件是中间语言的必要性以及大量数据的测试以保证结果具有统计意义。论文指出，在接下来的研究中，研究组致力于找出更简单且不依赖于中间语言的MR。

  文章指出新的MR，称之为$MR_{replace}$，在某些情况下，如果我们更改系统输入中相对独立的组件的值，则系统输出中仅仅有一小部分被改变（或不更改任何部分）。这和之前的**结构不变性**类似。

  输入微小的改变不应该对输出的整体结构产生影响。文章提出使用MRreplace来自动地测试英汉翻译系统。在我们对两个语言都很熟悉的前提下，我们可以设计生成测试用例的规则。
  之后举例论证短句子评估的重要性以及BLEU NIST指标对于短句子的表现性能不佳，在这里，文章指出不使用以上指标来评价短句子。

  实验设置

  测试对象是谷歌翻译和必应翻译，测试用例设计规则是采用主谓宾结构，动词采用likes和hates，主语与谓语基本不相关，选择美国前100受欢迎女性名字作为主语集合，100受欢迎品牌作为宾语集合，于是英语句子有20000句，分别获得谷歌和必应的API给出的翻译结果，共计40000对翻译结果。
  实验结果
  四万句分成四组，对比原始英语句差别只有一个名词的翻译对（两个原始句子两个各自的中文翻译），每组各自990000个翻译对，不一致性被判定在两个结果句子中的差别大于一处。
  实验结果表示对于谷歌或者必应，他们的like或者hate组别的不一致性结果相似，谷歌不一致性数据为13.88%和16.07%，必应的不一致性结果数据较好为3.50%和3.72%。

  局限性：没有考虑用同样是名词或者动词的名字影响，只关注于一致性的检测，测试集合数量较小，动词只有两种。
  尽管如此，可以解释开篇提出的研究问题。结果是谷歌必应翻译极短句子可能会失败或者具有潜在的不一致性，另外经验表明谷歌翻译比必应表现好，这与本次实验结果相反，由此认为此次不一致性测试的测试集合种类与以往不同。这也表明了评估一个翻译软件的质量应该从多元化的角度去评定。

  未来的工作方向，提出多元化的蜕变关系且分析不同种类的输入来做测试