尽可能少，尽可能多：基于对比条件反射的过翻译和欠翻译检测

Jannis Vamvas[1] and **Rico Sennrich**[1,2]
[1]Department of Computational Linguistics, University of Zurich
[2]School of Informatics, University of Edinburgh
{vamvas,sennrich}@cl.uzh.ch

Abstract

内容的省略和添加是神经机器翻译中的一个典型问题。本文提出一种利用现成的翻译模型来检测此类现象的方法。利用对比条件作用，在给定相应的源或目标序列的情况下，将翻译模型下完整序列的可能性与其部分的可能性进行比较。这样即使在没有参考译文的情况下，也可以准确地指出译文中多余的词和源文本中未翻译的词。该方法的准确性与需要自定义质量估计模型的监督方法相当。

1 简介

神经机器翻译(NMT)容易出现添加多余目标词或遗漏重要源内容等覆盖错误。以前检测此类错误的方法使用参考翻译(Yang et al., 2018)或者采用一个单独的质量估计(QE)模型，该模型在针对语言对(Tuan et al., 2021; Zhou et al., 2021)的合成数据上进行训练。

本文提出了一种基于假设推理的无参考算法。我们的前提是，如果翻译使用尽可能少的信息而使用尽可能多的信息来传达源序列，那么翻译就具有最佳的覆盖率。因此，额外的错误意味着包含较少信息的翻译将更好地传达源。相反，省略错误意味着译文更适合信息较少的源序列。

采用我们的对比条件反射方法(Vamvas and Sennrich, 2021)，我们使用NMT模型的概率分数来近似这种覆盖率的概念。为源序列和翻译创建解析树，并将它们的组成部分视为信息单元。遗漏错误是通过系统地从源序列中删除部分成分和估计以该部分源序列为条件的翻译概率来检测的。如果概率分数高于以完整源为条件的翻译，则删除的组成部分可能在翻译中没有对应项(图1)。通过交换源序列和目标序列，我们将相同的原理应用于加法错误的检测。

当将检测到的错误与片段级别上人工标注的覆盖错误进行比较(Freitag et al., 2021)时，所提出方法超过了在大量合成覆盖错误上训练的监督QE基线。人工评分人员发现，单词级别的精度对于省略比添加更高，英语-德语翻译有39%的预测错误是精确的，中文-英语翻译有20%的预测错误。错误的预测可能会发生，特别是在翻译与源具有不同语法的情况下。我们相信，当人类仍然处于循环中时，例如在后期编辑工作流程中，我们的算法可以是一个有用的帮助。

发布了代码和数据来重现发现，包括一个英德和汉英机器翻译合成覆盖率错误的大规模数据集。[1]

2 相关工作

NMT中的覆盖错误 在各种语言的人工评价研究中都观察到了目标词的添加和省略，其中省略是出现频率较高的错误类型(Castilho et al., 2017; Zheng et al., 2018)。它们作为典型的翻译问题包含在多维质量度量(MQM)框架(Lommel et al., 2014)中。添加被定义为准确性问题，当目标文本包含源文本中不存在的文本时，省略被定义为准确性问题，即翻译中缺少内容，但源文本中存在内容。[2]

Freitag et al. (2021) 使用MQM手工重新标注提交给WMT 2020新闻翻译任务(Barrault et al., 2020)的英德和汉英机器翻译。他们的发现证实了最先进的神经机器翻译系统仍然会错误地添加和省略目标词，而且省略的情况比添加的情况更频繁。类似的模式可以在英法机器翻译中找到，这些机器翻译已经用文档级QE共享任务(Specia et al., 2018; Fonseca et al., 2019; Specia et al., 2020)的细粒度MQM标签进行了注释。

检测和减少覆盖率错误 基于参考的方法包括测量n-gram与参考的重叠(Yang et al., 2018)和分析与源词的对齐(Kong et al., 2019)，这项工作侧重于无参考的覆盖错误检测。

[1]https://github.com/ZurichNLP/coverage-contrastive-conditioning

[2]过度翻译和欠翻译的术语也在文献中被使用。MQM将这些术语用于翻译过于具体或过于不具体的错误。

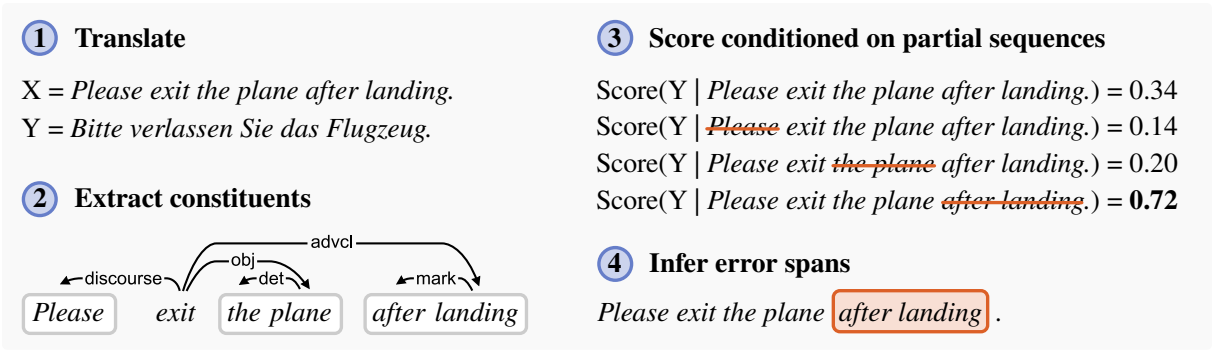

Figure 1: 如何检测遗漏错误的示例。德语翻译Y在落地后错误地离开(步骤1)。潜在的错误范围是从解析树中派生出来的(第2步)。像mBART50这样的NMT模型给以删除后的源为条件的Y分配的概率分数高于给以完整源为条件的Y的概率分数(步骤3)。这表明存在遗漏误差(步骤4)。

之前的工作采用了在标记的并行数据上训练的自定义QE模型。例如，Zhou et al. (2021)插入合成幻觉并训练Transformer来预测插入的span。类似地，Tuan et al. (2021)在合成噪声翻译上训练QE模型。本文提出了一种仅基于现有NMT模型的方法。

其他相关工作侧重于提高解码或培训期间的覆盖率，例如通过attention (Tu et al., 2016; Wu et al., 2016; Li et al., 2018;等等)。最近，Yang et al. (2019)发现对参考文献进行对比微调与合成遗漏可以减少NMT系统产生的覆盖错误。

3 方法

对比调节 翻译的属性可以通过根据对比源序列估计其概率来推断(Vamvas and Sennrich, 2021)。例如，如果在以反事实源序列为条件的NMT模型下，某种翻译更可能出现，则翻译可能是不充分的。

遗漏错误的应用 图1说明了对比条件反射如何直接应用于漏检错误的检测。我们通过系统地从源中删除组成部分来构造部分源序列。如果以这样的部分源为条件，翻译的概率分数(平均标记日志概率)更高，则将删除的成分视为翻译中缺失的成分。

来计算概率得分对于一个翻译Y给定一个源序列X，我们对每个目标token的对数概率求和，并将求和结果归一化为目标token的数量:

$$\text{score}(Y|X) = \frac{1}{|Y|}\sum_{i=0}^{|Y|}\log p_\theta(y_i|X, y_{<i})$$

添加错误的应用 我们将相同的方法应用于加法检测，但交换源语言和目标语言。即，我们使用一个NMT模型进行反向翻译，并以完整翻译和一组部分翻译为条件对源序列进行评分。[3]

潜在误差范围 在其最基本的形式中，除了标记化外，该算法不需要任何语言资源。对于包含n个标记的源语句，可以创建n个部分源序列，并删除第i个标记。然而，这种方法将依赖于一个激进的组合性假设，将所有代币视为独立的组成部分。

因此，本文建议从解析树中提取潜在的错误片段，特别是从通用依存分析器(de Marneffe et al., 2021)预测的依存树中，这些分析器广泛可用。这允许(a)跳过功能词，(b)在潜在错误范围集合中包括合理数量的多词范围。形式上，我们考虑满足下列条件的词跨度。

1. 潜在的错误跨度是依赖树的一个完整子树。
2. 它覆盖一个连续子序列。
3. 它包含了一个有趣的词性。

对于每个潜在的错误范围，我们通过从原始序列中删除该范围来创建一个部分序列。这仍然是一个简化的概念，因为一些部分序列将不符合语法。假设NMT模型可以产生可靠的概率估计，尽管输入不符合语法。

4 实验装置

在本节中，我们将描述用于实现和评估我们的方法的数据和工具。

评分模型 我们使用mBART50 (Tang et al., 2021)，这是一个序列到序列的Transformer，在许多语言的单语语料库上使用BART目标(Lewis et al., 2020; Liu et al., 2020)进行预训练，该目标在50种语言的以英语为中心的多语言MT上进行了微调。序列级概率得分通过平均计算所有目标标记的对数概率。如果英语是源语言，我们使用一对多的mBART50模型;如果英语是目标语言，我们使用多对一模型。

[3]另一种可能是保持翻译方向不变，并以源为条件对部分翻译进行评分。然而，部分翻译的不流畅可能会影响分数。

	Approach	Detection of additions			Detection of omissions		
		Precision	Recall	F1	Precision	Recall	F1
EN–DE	Supervised baseline	6.9±1.9	2.9±0.9	4.0±1.3	40.3±5.2	6.1±0.1	10.6±0.2
	Our approach	4.0	15.0	**6.3**	22.3	18.8	**20.4**
ZH–EN	Supervised baseline	4.3±0.6	4.7±0.7	**4.5**±0.6	49.6±0.6	9.4±1.0	15.9±1.4
	Our approach	1.7	40.6	3.4	25.8	62.0	**36.5**

Table 1: 在Freitag et al. (2021)上的gold数据集上的覆盖错误检测方法的片段级比较。对用不同随机种子训练的三个基线模型进行平均，报告标准差。

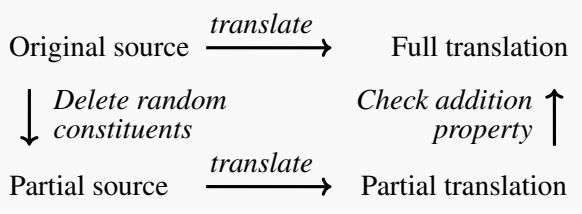

Figure 2: 为创建具有综合覆盖错误的机器翻译而设计的过程。完整翻译包含关于部分源的加法错误，而部分翻译包含关于原始源序列的省略错误。

误差范围　我们使用Stanza (Qi et al., 2020)进行依存分析，这是在来自通用依赖关系(de Marneffe et al., 2021)的数据上训练的各种语言的神经管道。利用通用词性标记(universal parts -of-speech tags, UPOS)来定义可能构成潜在错误范围的词性。具体来说，我们将常见名词、专有名词、主要动词、形容词、数字、副词和感叹词作为相关的词性。

金标准数据　我们使用最先进的英德、汉英机器译文进行评测，这些译文已经由Freitag et al. (2021)标注了翻译错误。[4] 我们将Online-*B*系统的翻译作为开发集，将其他系统的翻译作为测试集，不包括人工翻译。开发集用于识别上面段落中列出的覆盖错误范围的典型词性。

合成数据　还创建了合成覆盖误差，用于训练有监督的基线量化宽松系统。本文提出一种数据创建过程，受之前工作(Yang et al., 2019; Zhou et al., 2021; Tuan et al., 2021)的启发，但被定义为既适用于添加也适用于省略，并产生流利的翻译。

图2说明了这个过程。我们从原始源语句开始，通过删除随机选择的组成部分来创建部分源。具体来说，我们以15%的概率删除每个成分。然后对原始和部分来源进行机器翻译，得到完整和部分机器翻译。我们只保留完整的机器翻译不同于部分机器翻译的样本，并且可以通过加法来构建。

这使得我们可以将完整的翻译视为部分来源的过度翻译，而将添加的单词视为附加错误。相反，部分翻译被视为原始来源的欠翻译。通过将原始来源与完整译文、部分来源与部分译文进行配对来创建负例。[5]

我们的合成数据是基于WMT发布的单语新闻文本。[6] 为了训练基线系统，我们为每个语言对使用80k个唯一的源段。统计数据见表A3。

监督基线系统　遵循Moura et al. (2020)概述的方法，我们使用OpenKiwi框架(Kepler et al., 2019)在基于XLM-RoBERTa (Conneau et al., 2020)的基础上，为每个语言对训练一个单独的预测-估计器模型(Kim et al., 2017)。监督任务可以被描述为标记级二进制分类。每个标记都被分类为“OK”或“BAD”，类似于QE共享任务使用的单词级标签(Specia et al., 2020)。如果翻译中省略了源标记，那么源标记就是错误的;如果翻译中的标记是附加错误的一部分，那么翻译中的标记就是错误的。对于英语和德语，我们使用Moses tokenizer (Koehn et al., 2007)将文本分隔为带标签的标记;对于中文，我们在字符级别上标记文本。

在合适的地方，我们使用OpenKiwi的默认设置。我们微调了XLM-RoBERTa的大型版本，得到了一个与我们用于对比调节的mBART50模型相似的参数计数模型。我们训练了10个epoch, batch大小为32，并提前在验证集上停止。对于token分类，我们分别为源语言和目标语言训练了两个线性层(分别对应于省略和增加)。我们使用学习率为1e-5的AdamW (Loshchilov and Hutter, 2019)，在前1000步冻结预训练编码器。

[4] https://github.com/google/wmt-mqm-human-evaluation

[5] 请注意，合成数据集不包含同时带有添加错误和遗漏错误的翻译，这是一个限制。尽管如此，我们仍然希望在该数据集上训练的系统能够泛化这些示例，特别是在使用两个单独的分类器进行添加和省略的情况下。

[6] http://data.statmt.org/news-crawl/

		EN–DE	ZH–EN
Target	Addition errors	2.3	1.2
	Any errors	7.4	12.0
Source	Omission errors	36.3	13.8
	Any errors	39.4	19.5

Table 2: 人工评价:由我们的方法突出显示的span的单词级精度。

5 评价

5.1 与黄金数据的分段级比较

该方法的准确性可以通过Freitag et al. (2021)的人工评分来估计。

评价设计　我们使用MQM错误类型准确性/加法和准确性/省略，并忽略其他类型，如准确性/误译。如果任何一个人工评分者在段中任何地方标记了相同的错误类型，则将预测视为正确。[7] 我们从评估中排除部分可能注释不完整(因为评价者在标记了五个错误后就停止了)。为了便于实现，我们还排除了由多个句子组成的片段。

结果　金标准比较的结果如表1所示。该方法在检测两种语言对的遗漏错误方面明显超过了基线。然而，两种方法识别加法误差的准确率都不高，尤其是监督基线的召回率较低。考虑到它在合成测试集上的高性能(附录中的表A1)，似乎该模型不能很好地泛化到真实世界的覆盖率错误，突出了在纯合成数据上训练监督QE模型的挑战。

5.2 人工精度评价

进行了额外的单词级人工评估，以更详细地分析通过所提出方法获得的预测。我们的人工评分者被展示了在上述评估中被标记为真阳性或假阳性的片段，使我们能够量化单词级别的精度。

评价设计　我们为每个语言对雇佣了两名语言专家作为评价者。[8] 在两种类型的覆盖错误中，每个评分者都被展示了大约700个随机抽样的阳性预测。

评价者被展示了源序列、机器翻译和预测的误差范围。他们被问及高亮的跨度是否确实翻译得很糟糕，并被要求根据预定义的答案选项列表执行细粒度的分析(附录中的图3和4)。

[7]我们在本节中进行了段级别的评估，但没有量化单词级别的准确性，因为数据集不包含一致注释的覆盖错误片段。

[8]评价者每小时支付约30美元。

一部分样本由两位评分者标注。对于主要问题，双方的一致意见比较温和，英德语的科恩kappa值是0.54，中英语的科恩kappa值是0.45。对于更主观的后续问题的一致性较低(0.32 / 0.13)。

结果　细粒度的答案使我们能够量化我们的方法突出显示的span的单词级别的精度，包括特别的覆盖错误和一般的翻译错误(表2)。在检测英德翻译中的遗漏错误时，准确率比预期的要高，但在检测添加错误时仍然很低。详细答案的分布(附录中的图3和4)表明源语言和目标语言之间的语法差异导致了有关添加的误报。在附录F中提供了预测的示例，其中包括Freitag et al. (2021)的所有三个评分者都忽略了覆盖率错误的情况。

最后，表2显示许多预测的错误范围实际上是翻译错误，而不是狭义上的覆盖错误。例如，超过10%的中英译文被我们的评价者归类为不同类型的准确性错误，例如误译。

6 局限性和未来工作

我们希望自动检测覆盖错误可以对翻译和后期编辑有所帮助，因为手动检测这样的错误是很繁琐的。关于遗漏的结果是令人鼓舞的，建议进行用户研究，以验证预测对实践者的有用性。实际数据中样本较少的加法的检测还需要进一步的工作。单词级QE需要更高的准确性才能有所帮助(Shenoy et al., 2021)，因此在检测附加错误方面，基线和我们的方法的实际效用仍然有限。

还应该讨论推理时间。在附录C中，我们进行了比较，发现一对长句子的对比条件反射可能比基线前向传递花费多达10倍的时间。然而，这仍然只是最初生成翻译所需时间的一小部分。此外，限制所考虑的潜在误差范围可以进一步提高效率。

7 结论

我们提出了一种无参考的方法来自动检测翻译中的覆盖率错误。从对比条件作用衍生而来，所提出方法依赖于对部分序列的可能性的假设推理。由于任何现成的NMT模型都可以用于估计条件似然，因此不需要访问原始翻译系统或质量估计模型。在真实机器翻译上的评估表明，该方法在检测遗漏方面优于监督基线。未来的工作可以解决加法误差的低精度问题，这在我们用于评估的数据集中相对较少。

致谢

这项工作由瑞士国家科学基金会资助(项

目MUTAMUR;没有。176727)。我们要感谢Xin Sennrich促进了注释者的招聘，并感谢Chantal Amrhein和匿名评论者提供有用的反馈。

References

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Miceli Barone, and Maria Gialama. 2017. A comparative quality evaluation of PBSMT and NMT using professional translators. *16th Machine Translation Summit 2017*, pages 116–131.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568, Copenhagen, Denmark. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Xiang Kong, Zhaopeng Tu, Shuming Shi, Eduard Hovy, and Tong Zhang. 2019. Neural machine translation with adequacy-oriented learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6618–6625.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yanyang Li, Tong Xiao, Yinqiao Li, Qiang Wang, Changming Xu, and Jingbo Zhu. 2018. A simple and effective approach to coverage-aware neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 292–297, Melbourne, Australia. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

João Moura, Miguel Vera, Daan van Stigt, Fabio Kepler, and André F. T. Martins. 2020. IST-unbabel participation in the WMT20 quality estimation shared task.

In *Proceedings of the Fifth Conference on Machine Translation*, pages 1029–1036, Online. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Raksha Shenoy, Nico Herbig, Antonio Krüger, and Josef van Genabith. 2021. Investigating the helpfulness of word-level quality estimation for post-editing machine translation output. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10173–10185, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Belgium, Brussels. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Yi-Lin Tuan, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Francisco Guzmán, and Lucia Specia. 2021. Quality estimation without human-labeled data. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 619–625, Online. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2021. Contrastive conditioning for assessing disambiguation in MT: A case study of distilled bias. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Jing Yang, Biao Zhang, Yue Qin, Xiangwen Zhang, Qian Lin, and Jinsong Su. 2018. Otem&Utem: Over- and under-translation evaluation metric for NMT. In *Natural Language Processing and Chinese Computing*, pages 291–302, Cham. Springer International Publishing.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. 2019. Reducing word omission errors in neural machine translation: A contrastive learning approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6191–6196, Florence, Italy. Association for Computational Linguistics.

Zaixiang Zheng, Hao Zhou, Shujian Huang, Lili Mou, Xinyu Dai, Jiajun Chen, and Zhaopeng Tu. 2018. Modeling Past and Future for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 6:145–157.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

A 注释器指南

你将会看到一系列的原文和译文。文本中的一个或几个*span*被突出显示，并声称这些*span*被翻译得很糟糕。你被要求判断这个说法是否正确。高亮显示的*span*可以在源序列中，也可以在翻译中。如果源句子中有一个*span*，请检查它是否被正确翻译。如果翻译中有*span*，请检查它是否正确地传递了源。有时，多个*span*被高亮显示。在这种情况下，请将答案集中在翻译中最有问题的跨度上。第二步，要求你选择一个解释。一方面，如果你认为高亮的*span*翻译得不好，请选择你的解释来解释你的理由。另一方面，如果您不同意并认为该*span*翻译得很好，请首先选择一个解释为什么该*span*可能被标记为翻译不良。如果多种解释都同样合理，那么从最上面选择第一种解释。

	Detection of additions				Detection of omissions			
	Prec.	*Recall*	*F1*	*MCC*	*Prec.*	*Recall*	*F1*	*MCC*
EN–DE								
Supervised								
Baseline	98.8±0.4	98.0±.2	**98.4**±.2	**96.8**±.1	94.0±1.3	96.6±0.4	**95.3**±.5	**90.5**±.2
Ours	78.1	88.3	82.9	76.7	80.9	98.6	88.9	78.1
ZH–EN								
Supervised								
Baseline	87.2±1.5	75.7±.6	**81.0**±.3	**72.6**±.6	67.3±1.3	68.0±1.2	**67.7**±.9	**53.8**±.3
Ours	26.1	88.9	40.4	23.3	28.3	92.0	43.3	40.3

Table A1: 基于合成覆盖错误测试集的段级和词级*(MCC)*评估。

	Short sentence pair			Long sentence pair		
	Additions	Omissions	Both	Additions	Omissions	Both
Supervised baseline	-	-	25 ms	-	-	25 ms
Our approach	40 ms	45 ms	83 ms	165 ms	197 ms	365 ms
– excluding parser	18 ms	21 ms	38 ms	102 ms	144 ms	239 ms

Table A2: 在预测短句子和长句子对时的推理时间。由于我们没有使用针对效率进行优化的解析器，因此我们额外报告了推理时间，而不包括解析所需的时间。

B 综合误差评价

我们使用从合成数据中提取的测试分割来执行额外的评估。在片段级别，我们报告了精度，召回率和f1 -分数。就像在5.1部分，如果预测的覆盖错误在段的任何地方确实存在这种类型的覆盖错误，则预测在段级别上被视为正确。

在单词级别上，我们遵循之前在单词级别QE (Specia et al., 2020)上的工作，并报告测试集中所有标记的马修斯相关系数(MCC)。

结果　结果如表A1所示。监督基线在英德翻译上具有较高的准确性，在汉英翻译上具有中等的准确性。相比之下，所提出方法在合成误差上的表现明显比监督基线差。

C 推理时间

推断时间报告在表A2中。我们测量了在英语-德语的短句子对和长句子对上运行覆盖率错误检测方法所需的时间。短句子对取自图1，长句子对在源序列中有40个标记，在目标序列中有47个标记。我们在RTX 2080 Ti gpu上平均重复超过1000次。

所提出方法的较高推理时间可以用需要估计的翻译概率的数量来解释。在英德MQM数据集中，我们计算出平均每个句子30个分数，在汉英MQM数据集中，每个句子44个分数。不过，计算所有这些分数所需的时间只占生成翻译所需时间的一小部分(254 ms为短源句子，861 ms为长句子，假设波束大小为5)。

通过考虑更少的潜在错误跨度，可以减少所需的分数数量。此外，评分可以在多个翻译的批次中并行进行。最后，使用更有效的解析器或根本不使用解析器可以加快推理速度。

D 数据集统计

Dataset split	Number of segments			Number of tokens			
	Total	W/ addition	W/ omission	Src. OK	Src. BAD	Tgt. OK	Tgt. BAD
EN–DE Train	135269	18423	18423	2185918	58378	2197843	53911
EN–DE Dev	16984	2328	2328	273311	7398	275156	6781
EN–DE Test	16984	2328	2328	273277	7701	275036	7032
ZH–EN Train	110195	10697	10697	2576135	62311	1866567	37730
ZH–EN Dev	14149	1383	1383	326743	7562	236685	4244
ZH–EN Test	14026	1342	1342	322000	7566	234757	4882

Table A3: 4部分描述的合成覆盖错误数据集的统计信息。

Dataset split	Number of segments		
	Total	With an addition error	With an omission error
EN–DE Dev	1418	77	187
EN–DE Test	8508	407	1057
– without excluded segments	4839	162	484
ZH–EN Dev	1999	69	516
ZH–EN Test	13995	329	3360
– without excluded segments	8851	149	1569

Table A4: gold数据集的统计信息，请访问Freitag et al. (2021)。

E 合成覆盖错误的例子

英德例子

加法误差
部分资料来源:但他们还没有玩过。
完整机器翻译:Aber sie haben nicht gegen ein Team wie uns gespielt。

遗漏错误
完整来源:但他们还没有玩过against a team like us。
部分机器翻译:Aber sie haben nicht gespielt。

汉英示例

加法误差
部分来源: 医院和企业共同研发相关检测试剂盒，惠及更多患者。
全文翻译:医院和企业共同开发相关检测试剂盒，使更多cancer患者受益。

遗漏错误
完整来源: 医院和企业共同研发相关检测试剂盒，惠及更多肿瘤患者。
部分翻译:医院和企业共同开发相关检测试剂盒，使更多的患者受益。

F 对比条件反射预测的覆盖误差示例

英德例子

预测加法误差
消息来源:他补充道:“但现在他却事与愿违，这是一件可悲的事情。”
机器翻译:Er fügte **hinzu**:“Es ist jetzt auf ihn abgefeuert, aber das ist das Traurige。”
原始*MQM*评分*(Freitag et al., 2021)*:三位评分者没有标注相关的精度误差。
我们的人工评分者的回答:突出的目标跨度翻译得不错。它之所以突出显示，可能是因为它在语法上与源代码不同。
*highlight span*的含义:hinzu = '另外'

预测遗漏误差
来源:英国医疗**drug**供应仍不确定，无协议脱欧
机器翻译:Die medizinische Versorgung Großbritanniens ist im no - agreement - brexit noch ungewiss
原始*MQM*评分:三位评分者没有标注精度误差。
我们的人工评价者的回答:高亮的源跨度确实翻译得很糟糕。它包含的信息是在翻译中缺失，但可以推断或微不足道。

预测遗漏误差
消息来源:这家汽车制造商预计将在未来**few**天内公布其季度车辆交付情况。
机器翻译:Der Autohersteller wird voraussichtlich in den nächsten Tagen seine vierteljährlichen Fahrzeugauslieferungen melden。
原始的*MQM*评分:三个评分者没有标注相关的精度误差。
我们的人工评价者的回答:高亮的源跨度翻译得不错。*span*中的单词不需要翻译。

中英示例

预测加法误差
来源: 美方指责伊朗制造了该袭击，并对伊朗实施新制裁。
机器翻译:美国指责伊朗引发了这次袭击，并对其实施了新的制裁**on Iran**。
原始*MQM*评分*(Freitag et al., 2021)*:三位评分者没有标注相关的精度误差。
我们的人工评分者的回答:突出的目标跨度翻译得不错。没有发现可能导致预测的现象。

预测遗漏误差
来源: 目前已收到来自俄罗斯农业企业的约50项申请。
机器翻译:大约收到了50份来自俄罗斯农业企业的申请。
原始*MQM*评分:三位评分者没有标注精度误差。
我们的人工评价者的回答:高亮的源跨度确实翻译得很糟糕。它包含翻译中缺少的信息。
高亮*span*的含义: 目前 = '目前'

预测遗漏误差
来源: 他说，该系统目前在世界上有很大需求，但俄罗斯军队也需要它，
其中包括在北极地区。
机器翻译:他说，该系统目前在世界上需求量很大，但俄罗斯军队也需要它，包括在北极地区。
原始*MQM*评分:三位评分者没有标注精度误差。
我们的人工评价者的回答:高亮的源跨度翻译得不错。*span*中的单词不需要翻译。
高亮*span*的含义: 其中 = ' among'

G 详细的人工评价结果

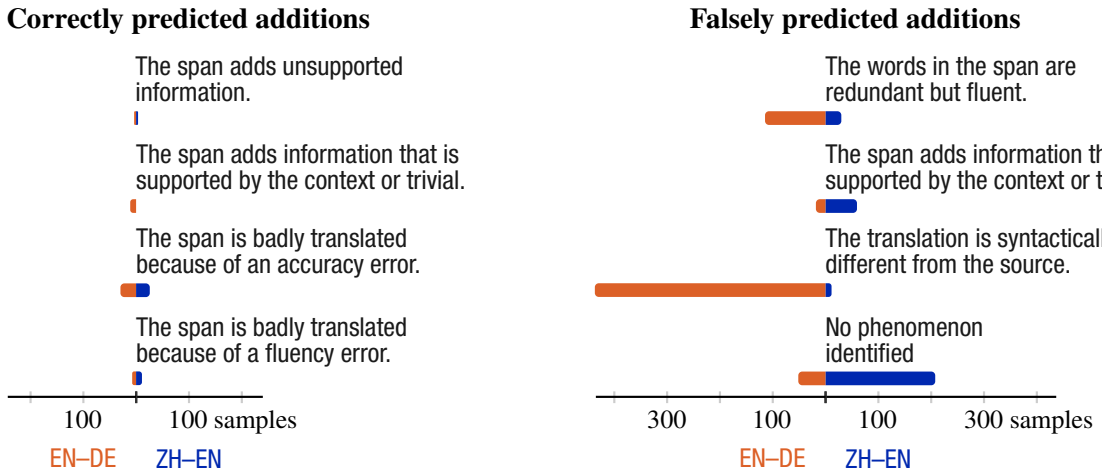

Figure 3: 人工评价预测加法误差的结果。如果人工评分者回答翻译中突出显示的跨度确实翻译得很糟糕，他们会得到左边的四个解释选项。否则，他们从右边的四个选项中选择。

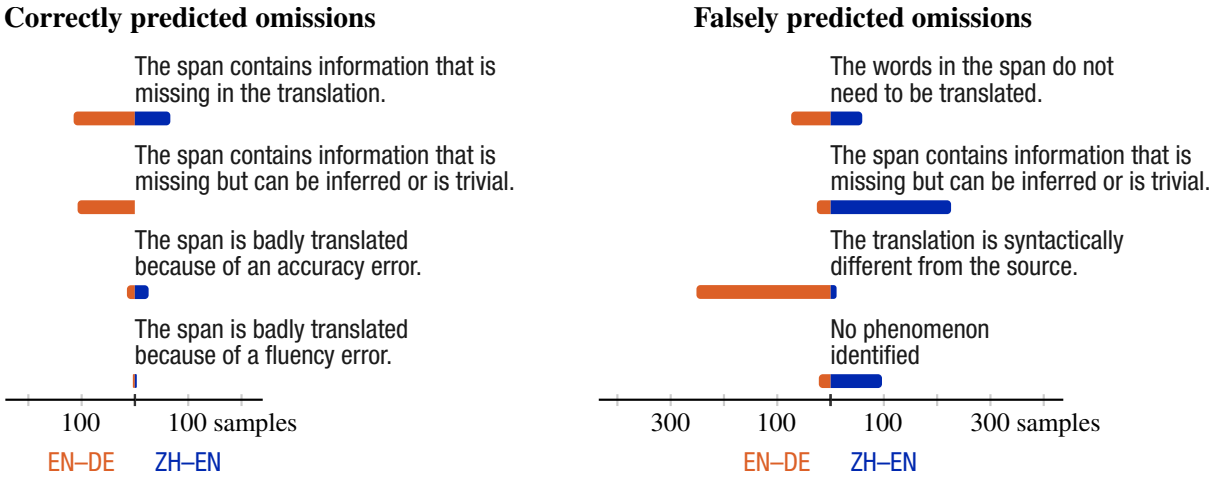

Figure 4: 预测遗漏错误的人工评价结果。如果人工评价者回答源序列中突出显示的跨度确实翻译得不好，他们将得到左边的四个解释选项。否则，他们从右边的四个选项中选择。