**Improving Disfluency Detection by Self-Training a Self-Attentive Model**

会议：ACL2020

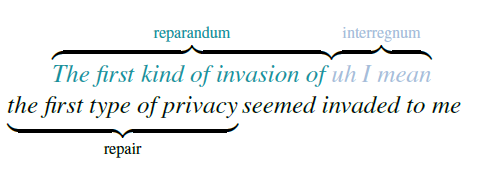
代码：<https://github.com/pariajm/joint-disfluency-detector-and-parser>

摘要

使用语境化词嵌入（例如ELMo或BERT）的自注意神经句法解析器目前在联合分析和语音记录中的语气检测方面产生了最新的成果。 由于上下文化词的嵌入在大量未标记的数据上进行了预训练，因此使用其他未标记的数据来训练神经模型似乎是多余的。 但是，我们显示出自我训练（一种用于合并未标记数据的半监督技术）为自注意解析器在流失性检测方面设置了新的最新技术，证明了自我训练提供了与预训练正交的优势。 训练的上下文化单词表示形式。我们还显示，集合自训练的解析器可为流失检测提供进一步的收益。

1简介

语音引入了书面文字中未出现的挑战，例如存在不满情绪。不满是指正常语音流中的任何中断，包括错误的开始，更正，重复和充满的停顿。 Shriberg（1994）定义了言语疏散的三个不同部分，分别称为reparandum，regregum和repair。 如下例所示，reparandum的The first kind of invasion of是发声的一部分，该部分被替换或修复，间断语uh I mean（由充满的停顿uh和I mean话语标记组成）是可选的不流利部分，而修复的the first type of privacy取代了reparandum。 流利的版本是通过删除reparandum和interregnum得到的。



本文将重点讨论联合不流利检测和转录语音的选区解析。 在Switchboard树库语料库（Godfrey和Holliman，1993； Marcus等人，1999）中，这是用于分析会话语音研究的标准语料库，reparanda，填充的停顿和话语标记主要由EDITED，INTJ和PRN节点控制， 分别（见图1）。 填充的停顿和话语标记属于单词和短语的有限集合，因此INTJ和PRN节点很难检测到（Johnson和Charniak，2004）。 然而，检测已编辑的节点是具有挑战性的，并且是不满检测模型的主要焦点。

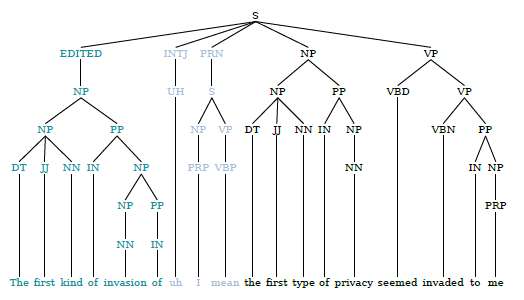


图1：来自Switchboard语料库的解析树，其中reparandum The first kind of invasion of、filled pause uh和discourse marker I mean以edited、intj和prn节点为主。

Jamshid Lou等人(2019)表明，一个自注意的选区解析器在联合解析和disfluency检测方面取得了最先进的结果。他们观察到，由于Switchboard树既包括句法选区节点，也包括表示disfluency的edited节点，因此训练一个解析器来预测Switchboard树可以被视为多任务学习（其中任务是句法解析和识别disfluency）。在本文中，我们扩展了Jamshid Lou等人(2019)中的多任务学习，以探索自我训练(McClosky等人，2006)和合集(Kitaev等人，2019)对自我注意力解析器性能的影响。我们旨在回答关于最先进的自注意力解析器的两个问题：

自训练是否能提高自注意力解析器在disfluency检测上的性能？自训练是一种半监督技术，用于将未标记的数据纳入新的模型中，即使用在手动标记（即黄金）数据上训练的现有模型来标记未标记的数据。自动（即白银）标记的数据被视为真理，并与黄金标记的数据相结合，重新训练一个新的模型（McClosky等人，2006；Choe和Charniak，2016）。由于神经模型使用丰富的语言表征，在大量未标记的数据上预先训练（Peters等人，2018年；Devlin等人，2019年），我们可能会期望自我训练不会给自我注意的解析器增加新的信息。然而，令人惊讶的是，我们发现自我训练提高了基于BERT的自我注意力解析器的disfluency检测f-score，证明自我训练提供了与预先训练的上下文嵌入正交的好处。

Ensembling是否能改善语音转录中的歧义检测？合集是一种常用的改进解析的技术，其中在推理时将在相同或不同数据上训练的同一模型的多个实例的分数结合起来（Dyer等人，2016；Fried等人，2017；Kitaev等人，2019）。我们期望合集解析器也能提高模型在歧义检测上的性能。我们展示了合集四个自训练的解析器（使用不同的 BERT词表征）通过平均它们的跨度标签得分，与单一的自训练解析器相比，增加了disfluency检测f-score。

**2相关工作**

对于常规句法解析器而言，解析语音记录本是一项挑战，主要是因为存在歧义。 在流利的句子中，reparandum和repair之间的关系与句子中的其他词不同。 repair通常是reparandum的“粗糙副本”，使用大致相同的单词顺序使用相同或相似的单词（Charniak and Johnson，2001）。 旨在捕获树状结构的常规语法分析器无法检测到“粗糙副本”，而粗糙副本是流利性的有力指标。 而且，reparandum和repair通常不形成语法短语，这使得检测reparandum更加困难。 由于这些原因，开发了专门的流量检测模型以在解析之前消除流量（Charniak和Johnson，2001； Kahn等，2005； Lease和Johnson，2006），或者在解析器中添加了特殊机制来处理流量（Rasooli和Tetreault）。 ，2013； Honnibal和Johnson，2014； Yoshikawa等，2016）。 基于常规解析的模型可以将流离失所的句法位置用作reranker中的功能（Johnson等，2004）。 通过训练联合解析和不满检测模型，可以在神经模型中获得类似的收益。 在这种多任务学习设置中，句法信息有助于神经模型更准确地检测出语气差异（Jamshid Lou等人，2019）。

对于传统的句法解析器来说，解析语音记录本是一个挑战，主要是由于不通顺句的存在。在不通顺的句子中，reparandum和repair的关系与句子中的其他词不同。repair通常是对reparandum的 "粗略复制"，使用相同或相似的词，词序大致相同（Charniak和Johnson，2001）。传统的句法解析器是为了捕捉树状结构而设计的，但却无法检测到 "粗略的副本"，而这正是不连贯性的强烈指标。此外，reparandum和repair的通常不形成一个句法短语，这使得检测reparandum更加困难。基于这些原因，人们开发了专门的歧义检测模型，以便在解析之前去除disfluency（Charniak和Johnson，2001；Kahn等人，2005；Lease和Johnson，2006），或者在解析器中添加特殊机制来处理disfluency（Rasooli和Tetreault，2013；Honnibal和Johnson，2014；Yoshikawa等人，2016）。传统的基于解析的模型可以将disfluency的句法位置作为reranker中的一个特征（Johnson等人，2004）。在神经模型中，可以通过训练一个联合解析和disfluency检测模型来获得类似的收益。在这种多任务学习环境中，句法信息有助于神经模型更准确地检测disfluency（Jamshid Lou等，2019）。

对于使用上下文嵌入（例如ELMo和BERT）的Transformer模型，已经报道了最新的disfluency检测结果（Jamshid Lou等人，2019年; Tran等人，2019年; Dong等人，2019年）。 Transformer的自我关注机制对于捕获单词之间的“粗糙副本”依存关系显然是有效的。 最近的一项研究表明，与纯文本模型相比，prosody会稍微提高自注意模型的解析性能，尤其是在长句子中（Tran等人，2019）。 在本文中，我们使用自注意模型进行联合disfluency检测和选区解析。

失语检测模型通常在Switchboard语料库上进行训练和评估。Switchboard是最大的语义disfluency标注数据集。然而，Switchboard中只有5.9%的词是disfluency的（Charniak和Johnson，2001）。

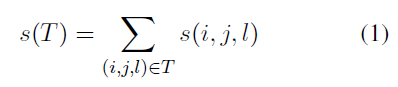
为了减轻标记数据的稀缺性，一些研究通过以下方式利用了其他数据：（i）在大量未标记数据上经过预训练的情境化嵌入（Jamshid Lou等人，2019; Tran等人，2019; Bach和Huang （2019年）和（ii）通过以disfluency的形式向流利的句子中添加噪音（例如在句子中重复，删除或插入单词）而生成的合成数据（Wang等人，2018； Bach and Huang，2019； Dong等人，2019）。 相比之下，本文着重于自我训练，这是一种简单的半监督技术，已在包括解析在内的各种NLP任务中有效（McClosky等人，2006; Clark等人，2018; Droganova等人，2018） 。 据我们所知，这是研究自我训练神经disfluency检测模型的第一项工作。

通常用于改进解析的另一种技术是集成。 整合是一种模型组合方法，其中将多个模型的分数（它们可以是相同或不同的模型，在相同或不同的数据上训练，具有不同的随机初始化）以某种方式进行组合（Dyer等人，2016； Choe和 Charniak，2016; Fried等人，2017）。 解析书面文本的最新技术是四个基于BERT的自注意解析器的集合，其中解析器通过平均其跨度标签得分来合并（Kitaev等人，2019）。 尽管集成在解析中被广泛使用，但尚未对其进行disfluency检测进行研究。 在本文中，我们还探讨了组装几个基于解析的disfluency检测模型对disfluency检测性能的影响。

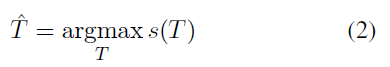
**3 Model**

继Jamshid Lou等(2019)之后，我们使用一个自我关注的选区解析器进行联合disfluency检测和句法解析2。该解析模型基于Kitaev和Klein（2018）引入的架构，该架构在以下方面是最先进的：（i）解析书面文本（Kitaev等人，2019；Fried等人，2019），（ii）解析转录语音（Tran等人，2019），以及（iii）联合解析和disfluency检测（Jamshid Lou等人，2019）。

自注意力解析器通过计算其标记的成分跨度上的potentials总和，给每棵树T分配一个分数s(T)。



其中是从字符串位置i开始并以标签l结束于位置j的成分的分数。 解析器的输入是与句子中单词序列相对应的向量序列，其后是一个或多个自我注意层。 对于每个跨度（i; j），通过减去跨度的开始和结束的表示来构造隐藏向量。 跨度分类器包括两个完全连接的层，后跟一个非线性度，为每个跨度分配标签分数。 然后，为给定句子找到最高得分的解析树，如下所示：



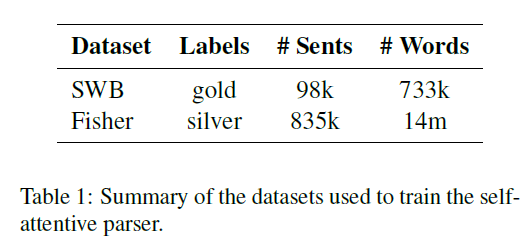
使用修改后的CYK算法。 Kitaev和Klein（2018）中引入的解析器依靠外部POS标记器来预测终端标签，但是由于当不使用外部POS标记器时解析器的准确性不会降低，因此我们在这里使用其解析器时无需使用外部POS标记器（因此， 所有的终端标签都是UNK）。有关更多详细信息，请参见Kitaev和Klein（2018）。

**3.1上下文化嵌入**

我们通过在训练过程中对参数进行微调，将BERT（Devlin等人，2019）整合到我们的自注意解析器中。 继Kitaev等（2019），我们在BERT的输出上应用了一个学习的投影矩阵，以将矢量投影到我们想要的尺寸。 然后将表示形式输入解析器。 BERT学习子词单位的表示形式，因此要提取单词表示形式，我们考虑句子中每个单词的最后一个子词单位的表示形式（Kitaev等，2019）。

3.2自我训练

我们在Penn Treebank-3 Switchboard语料库上训练自我专注的解析器，该语料库包含带有gold disfluency标记的解析树（Godfrey和Holliman，1993； Marcus等，1999）。 使用训练后的模型，我们解析未标记的数据，并将silver的解析树添加到gold Switchboard训练数据中，并使用扩大的训练集来重新训练自我专注的解析器。 我们使用的未标记数据包括Fisher Speech Transcripts Part 1和第二部分（Cieri等，2005）。 表1总结了用于训练自注意解析器的不同数据集。



4实验

根据Charniak和Johnson（2001），我们将Switchboard分为训练集，开发集和测试集，如下所示：训练数据由sw [23] .mrg文件组成，开发数据由sw4 [5-9]组成 .mrg文件和测试数据由sw4 [0-1] .mrg文件组成。所有不完整的词和标点符号均已从数据中删除，因为它们在实际的ASR应用程序中不可用（Johnson和Charniak，2004年）。

**4.1基准线**

基准线是在具有BERT字词表示法的Gold Switchboard语料库上训练的自我专注解析器。 基于BERT的解析器是当前的最新技术，为我们的工作提供了非常强大的基础。 我们使用四种不同的BERT模型（即BERTBASE [casedjuncased]和BERTLARGE [casedjuncased]）训练了不同版本的基线解析器，然后在Switchboard开发集中选择了最佳模型，即BERTBASE [cased]。 我们还通过优化分析EDITED节点F（SE）的性能来调整超参数。 在Switchboard开发集上的初步实验表明，Kitaev等人（2019）给出的超参数表现良好; 因此，这就是我们在这里使用的。 由于随机种子导致不同的结果，因此在本文中，我们报告了使用不同的随机种子初始化的每个模型的5次运行的平均得分。

4.2评估指标

我们根据解析的准确性以及流水线检测来评估自我专注的解析器。由于某些词在解析树中被标识为“已编辑”，因此我们可以衡量解析器将单词分类为“已编辑”的性能。我们还可以评估解析器识别所有disfluency单词的准确度，即由EDITED，INTJ或PRN节点主导的单词。 因此，我们报告了成分跨度（S）和单词位置（W）的精度（P），召回率（R）和f分数（F），其中每个单词位置都被包含该单词的所有成分视为标记。 我们还报告了组成跨度和词位置的子集的结果：（i），标记为EDITED的组成跨度的集合，（ii），由一个或多个EDITED节点控制的词位置的集合，以及（iii） ，由一个或多个EDITED，INTJ或PRN节点控制的一组单词位置。 有关更多详细信息，请参见Jamshid Lou等 （2019）。

**4.3不同数量的**Silver**训练数据**

为了找到额外的Silver训练数据的最佳比例，我们在每个迷你批次的Silver分析树中选择n％（范围从10％到90％）的训练数据，其余的从gold树中选择。 这具有与称重主要金语料库相同的效果，如McClosky等人（2006）所述。 图2中显示了使用不同比例的Silver解析树的结果。基于BERT的解析器是我们的最佳模型，其中包含40％的silver Fisher trees和60％的gold Switchboard树。 换句话说，对于批量大小为30的批次，每个微型批次中的12个解析树来自silver的Fisher数据，而18个解析树来自gold Switchboard。 本文所有的自训练结果均使用此比例的gold and silver解析树。

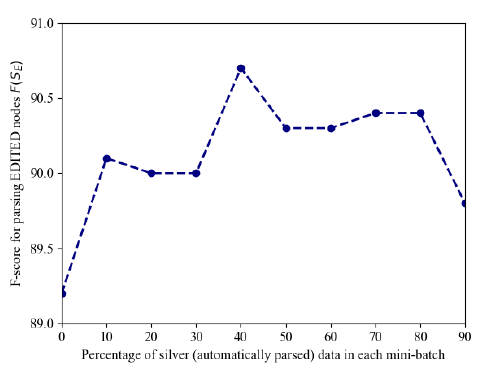
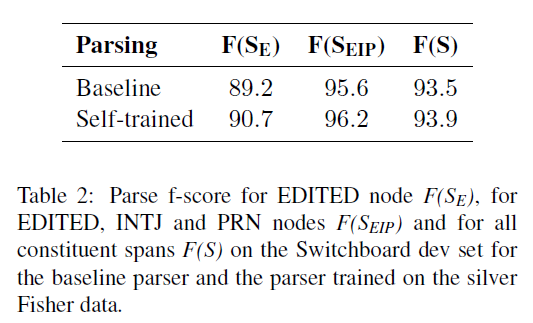


图2：基于BERT的自注意解析器的EDITED点f分数F（SE），其为来自银色费舍尔树的每个微型批处理中训练数据百分比的函数。

**4.4自我训练会提高自我专注的解析器的性能吗？**

表2和表3在解析和disfluency检测方面比较了基线解析器和自训练解析器。 在Silver Fisher数据上经过自训练的解析器提高了解析和disfluency检测性能，表明基于BERT的模型受益于附加的带有银标签的数据。自我训练对于识别EDITED的disfluency节点特别有效（f得分提高了1：5％）。 Switchboard中只有5.9％的单词是disfluent,，并且BERT仅在流利的文本（例如书籍和Wikipedia）上受过训练，因此基线解析器可能会缺少流利的培训示例。 结果，在会话语音语料库上的自我训练可以弥补流失的黄金数据的不足。 为了探索这一点，我们尝试了对各种流畅的干净数据集进行自我训练，其中包括Gigaword 5（未标记新闻专线语料库）以及WSJ和Brown（包括书面文字的黄金解析树），但是性能并未显着提高。 这表明，解析器从其他域内（即会话性）白银数据中受益，而不是其他域外（即书面）的银/金数据中。 此外，如果我们在训练过程中学习嵌入而不是使用预训练的BERT，则EDITED词f得分将从90.9％下降到86.4％，而对Fisher进行自我训练的结果几乎没有改善（EDITED词f增加0.2％ 得分，而使用BERT则提高1.5％）。 这表明当基线模型的功能足以预测准确的银标签时，自训练效果很好。



为了进一步研究自我训练对disfluency检测的影响，我们从Switchboard dev集中随机选取100个包含disfluency的句子。

我们根据Shriberg(1994)的语音修复类型学将 disfluencies分为重复、纠正和重新开始。重复是指disfluency的重复部分和修复部分相同的修复，而修正是指重复部分和修复部分不同的修复（这就更难发现了）。重新开始是指说话人放弃一个句子，开始一个新的句子（即修理是空的）。如表4所示，自训练的解析器在检测所有类型的不连贯性方面都优于基线。特别是在检测修正和重新开始方面有更好的表现，与重复相比，这些类型的语义不通畅更具有挑战性。

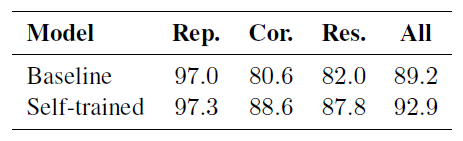
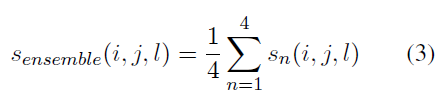


表4：在包含158个不同结构的Switchboard开发集的子集上，针对不同类型disfluency的EDITED单词f分数F（WE）-包括90次重复（Rep。），54次更正（Cor。）和14次重启（Res。 ）。

4.5集成解析器是否可以改善disfluency检测？

我们研究合奏对自我专注分析器性能的影响，其中我们通过对分析器的跨度标签得分进行平均来组合分析器，如下所示：



我们尝试了不同的解析器的合集，当我们使用四种BERT词表征训练基线解析器四次时，取得了最好的效果，即BERTBASE [casedjuncased]和BERTLARGE [casedjuncased]，并在推理时合并结果（Kitaev等，2019）。 集成的模型不仅反映了不同的预训练表示形式的变化，而且还反映了模型初始化时的随机性。 如表6所示，集合和自训练都可以提高基线单一模型在解析和检测EDITED disfluency节点时的性能。 自训练比集合更有效，尤其是对于EDITED节点检测。对于5个随机重启中的每一个不同的BERT模型，结合了最佳的自训练解析器而言，据报告获得了最佳结果。

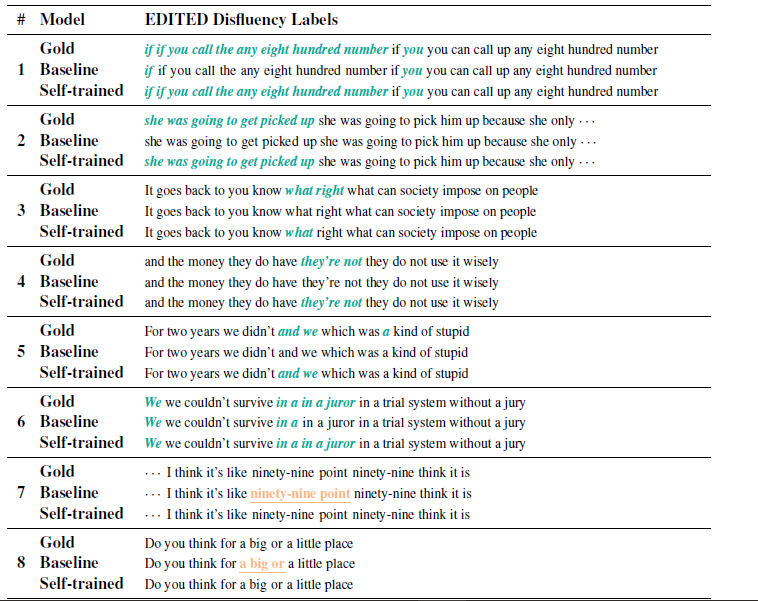


表5：来自Switchboard开发人员集合的一些示例以及由基线和最佳自训练解析器给出的相应EDITED disfluency标签以及gold（即正确的）标签。 绿色（和斜体）词表示正确标记的流利单词，橙色（和带下划线的词）表示流利的单词，这些单词被错误地标记为disfluencies.。

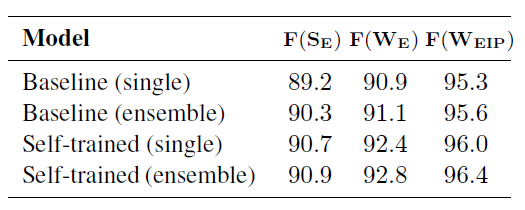


表6：针对Switchboard开发集上的不同模型，分析EDITED节点F（SE），EDITED字词F分数F（WE）以及EDITED，INTJ和PRN字词F分数F（WEIP）的f分数。 “ single” =单个解析器，“ ensemble” = 4个解析器的集合。

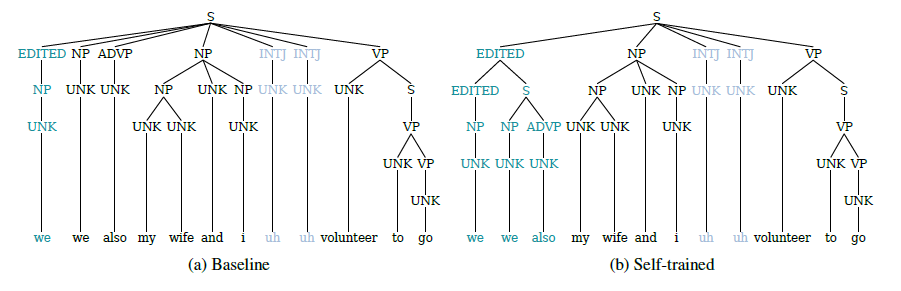


图3：来自Switchboard开发人员集的句子由基线模型（左）和自训练模型（右）解析。 自训练模型得到的解析树与黄金解析树相同。

**5结果**

我们将最佳模型的性能与之前在Switchboard测试集上的工作进行了比较。 如表7所示，我们的模型优于以前的解析工作。 我们模型的解析结果高于Tran等人。（2018），它利用了韵律提示以及基于文本的功能。

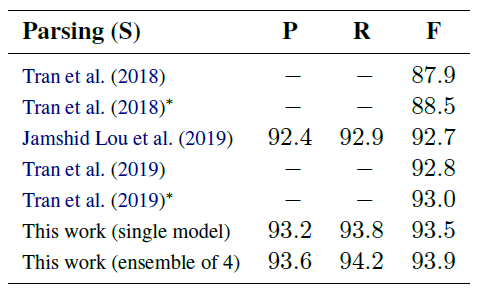


表7：解析总机测试集上所有组成范围的精度P，调用R和f分数F。文字+韵律模型。 P ＝ P（S），R ＝ R（S）和F ＝ F（S）。

我们将自注意解析器的性能与最新的不相容性检测模型进行比较。 如表8所示，我们的模型具有最佳的f分数。 我们还将模型与先前报告EDITED，INTJ和PRN字f分数进行流失检测的工作进行了比较，发现我们的模型具有最佳性能（请参见表9）。 与王等人相比。 （2018）使用GAN来利用其他未标记的数据，而Bach and Huang（2019）使用了综合数据，我们的模型显着提高了召回率。 这表明标准技术（例如自我训练和合奏）与这些专门的复杂方法一样好或更好。

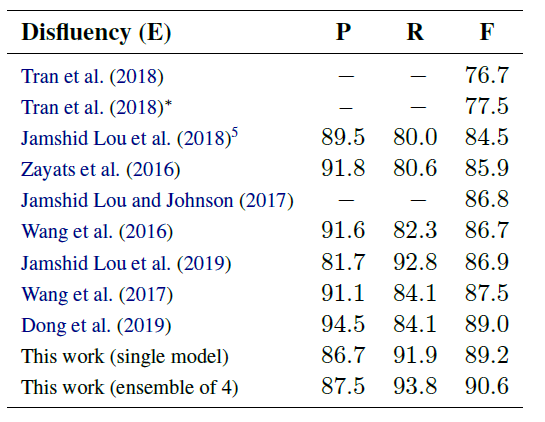
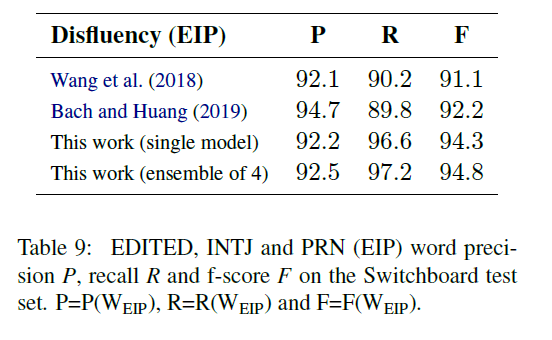


表8：EDITED单词精度P，在配电盘测试仪上调用R和fscoreF。 文字+韵律模型。 P = P（WE），R = R（WE）和F = F（WE）。



**5.1定性结果**

我们对Switchboard开发人员集进行定性分析，以描述基线模型无法检测到，但经过自我训练的模型可以检测到。 我们在表5中提供了具有代表性的示例。通常，自训练模型更擅长检测长复数校正（＃1-4），重新启动（＃5）和类似口吃的重复（＃6）。 它还能更好地区分流利的重复和流利的并行结构与流利的重复和校正类型（＃7和8）。 图3描绘了一个由基线和自我训练的自我专注解析器解析的句子，其中自我训练的模型正确地预测了所有不满编辑节点。 如第3节所述，我们不使用外部POS标记器，因此从原始文本进行解析时POS标记不可用。 这就是为什么图3中的所有预终端标签都用虚拟令牌（即UNK）显示的原因。

**6总结**

我们介绍了一种新的最新技术，可用于转录语音的联合disfluency检测和选区解析。 我们表明，自我训练和整合是改善disfluency检测的有效方法。 对结果的定性分析还表明，自我训练有助于检测出各种类型的出气，包括纠正和重启。 在未来的工作中，我们打算探索自我培训的思想来解析书面文本。 我们还旨在将语法分析和自我训练与自动语音识别更加紧密地集成在一起。 第一步是开发用于解析ASR输出而不是语音记录的解析模型。