(2021 ACL) Exploiting Document Structures and Cluster Consistencies for Event Coreference Resolution

时间：2021/8/4

1. **先前工作的两个限制**

* 限制一

捕获事件提及之间的文档级上下文(即，超出事件提及的两个句子)可能会为ECR提供有用的信息。先前用于ECR的深度学习模型仅试图通过手工设计的特征对文档级上下文进行编码，仍然遭受特征稀疏问题；此外，这种先前的工作无法利用文档中与ECR相关的对象(例如，实体提及、上下文词)及其连接/交互(可能超出句子边界)来帮助表征学习。

受这个问题的启发，作者提议为文档形成图形(称为文档结构)，以明确捕获相关对象和交互，用于学习事件提及的表示向量。上下文词、实体提及和事件提及将作为我们文档结构中的节点，因为它们与ECR有直观的相关性。然后将利用不同类型的知识源来连接文档结构的节点，特色的话语信息(例如，连接共有的实体提及)、句法信息(例如，直接链接事件提及及其自变量)和语义相似性(例如，连接具有相似含义的单词/事件提及)。这种丰富的文档结构允许我们在文档级上下文的句子级之外为ECR建模相关对象的交互。

* 限制二

现有的用于ECR的深度学习模型未能利用黄金聚类(由人类提供)和预测聚类(由模型生成)之间的一致性来促进表征学习。

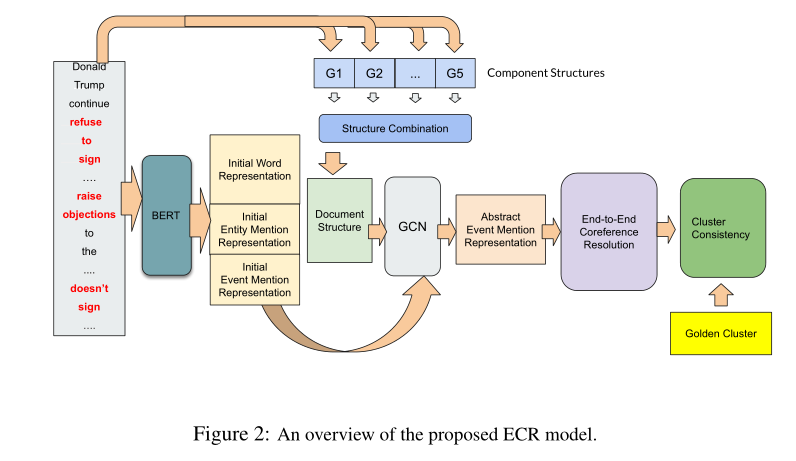
直观的是，如果ECR模型的预测事件聚类与数据中的黄金事件聚类更相似，则ECR模型可以获得更好的性能。为此，作者建议在黄金聚类和预测聚类之间获得不同的不一致性度量，这些度量将被合并到整体损失函数中以实现最小化。因此，我们期望两种类型的聚类之间的一致性/相似性正则化可以提供有用的训练信号来改善事件记录中事件提及的表示向量。

1. **模型**

在ECR中，给定一个输入文档D = {w1，w2，…，wN}(由N个单词/标记组成)，带有一组事件提及E = {e1，e2，…，e|E|}，目标是将E中的事件提及分组到聚类中，以捕获提及之间的共指关系。

ECR模型由四个主要部分组成:

1. 文档编码器，将单词转换为表示向量；
2. 文档结构，为文档创建图并学习事件提及的丰富表示向量；
3. 端到端共指消解，同时解决D中实体提及的共指；
4. 聚类一致性正则化，基于黄金和预测事件提及聚类之间的一致性约束来正则化表示向量。



* 1. **文档编码器（Document Encoder）**

通过将D馈送到预先训练的语言模型BERT中，将每个单词wi∈D转换成一个表示向量。为了用BERT处理长文档，将D分成有512个连续单词片段（word-peices）的片段，这些片段将被单独编码。得到的序列X = x1，x2，… ，xn被发送到下一步进行进一步计算。

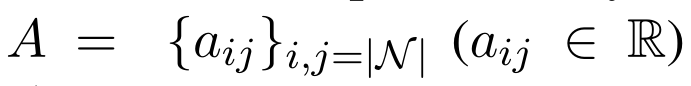
\*wordPiece字面理解是把word拆成piece一片一片。WordPiece的一种主要的实现方式叫做BPE（Byte-Pair Encoding）双字节编码。BPE的过程可以理解为把一个单词再拆分，使得我们的此表会变得精简，并且寓意更加清晰。比如"loved","loving","loves"这三个单词。其实本身的语义都是“爱”的意思，但是如果我们以单词为单位，那它们就算不一样的词，在英语中不同后缀的词非常的多，就会使得词表变的很大，训练速度变慢，训练的效果也不是太好。BPE算法通过训练，能够把上面的3个单词拆分成"lov","ed","ing","es"几部分，这样可以把词的本身的意思和时态分开，有效的减少了词表的数量。

* 1. **文档结构（Document Structure）**

该组件旨在使用交互图G = {N，E}来学习事件提及的表示向量，该交互图有助于丰富事件提及与相关对象的表示向量以及文档级别的交互。

**节点：N = D ∪ E ∪ M**（D是单词/标记集，E是事件提及集，M是实体提及集**）**

**边：**边将由邻接矩阵A 表示



为N中的两个节点ni和nj计算一个值/得分aij，用来估计nj对于ni的表示计算的重要性(或交互水平)。作者探索了三种类型的信息来设计模型中G的边E(或计算交互分数aij)，包括基于话语（discourse-based）、基于语法（syntax-based）和基于语义（semantic-based）的信息。

* **Discourse-based Edges**

作者提出利用三种类型的话语信息（discourse information）来获得交互图G，即句子边界（sentence boundary）、共指结构（coreference structure）和文档D中事件/实体提及的提及文段（mention span）。

* + Sentence Boundary

作者获取这个信息的动机是，出现在相同句子中的事件/实体提及往往比出现在不同句子中的事件/实体提及在上下文中更相关。

为N中的节点ni和nj计算基于句子边界的交互得分aijsent，其中如果ni和nj是D中相同句子的事件/实体提及(即ni，nj∈ E ∪ M)，则aijsent = 1；否则为0。

* + Entity Coreference Structure

共指结构关注句子之间实体提及的连接，用共指结构的上下文信息来丰富它们的表达。

为每对节点ni和nj计算基于共指的分数aijcoref，如果ni和nj是D中的共指实体，aijcoref设置为1，否则为0。

* + Mention Span

为了将事件和实体提及与上下文单词连接起来进行表示学习，作者使用基于提及文段的交互评分。

如果ni是实体/事件提及nj (nj∈ E∪M)的文段中的单词(ni∈ D)，则aijspan仅被设置为1(即，否则为0)，反之亦然。

* **Syntax-based Edges**

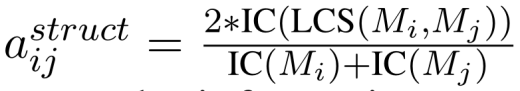
我们期望D中句子的依赖树能够为连接N中的节点提供有益的信息，以便在ECR中进行有效的表征学习。作者建议使用D中单词之间的依赖关系/连接来获得N中每对节点ni和nj的基于语法的交互得分，通过继承D中句子的依赖树的图结构，如果ni和nj是同一个句子中的两个单词(即ni，nj∈ D)，并且在对应的依赖树中它们之间有一条边，则将aijdep设置为1，否则为0。

* **Semantic-based Edges**

作者的动机是，如果ni在语义上与nj更相关，则节点ni将对ECR的另一个节点nj 的表示计算贡献更大。

作者探索一种新的语义信息源，该信息源依赖于单词的外部知识来计算ECR文档结构中节点N之间的交互分数。期望单词的外部知识为D中的上下文信息提供补充信息，从而进一步丰富N中节点的整体交互得分。作者利用单词网络(WordNet)来获得D中单词的外部知识。

* 首先，使用词义消歧(Word Sense Disambiguation，WSD)工具将每个单词节点ni∈D映射到WordNet中的synset节点Mi；
* 然后，使用它们在WordNet中链接的synsets节点 Mi和Mj的基于结构的相似性，为D中的每对单词节点ni和nj计算基于知识的相似性得分。



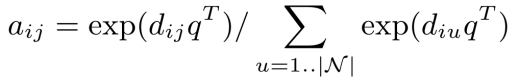
IC和LCS分别代表了synset节点的信息内容和WordNet层次结构中两个synset的最小公共子集（the least common subsumer）。

**结构组合（Structure Combination）**

现在，我们已经为每对节点生成了五个分数，将每个节点对ni和nj的五个分数分组为维度为5的向量dij

IMG_256

由此，可以计算ni和nj的总体交互分数aij



q是维度为5的可学习向量。

**表示学习（Representation Learning）**

给定具有邻接矩阵A = {aij}i，j=|N|的组合交互图G，作者使用GCNs来归纳ECR的N中节点的表示向量。

GCN模型将节点ni∈ N的初始表示向量作为输入，将vi组织成输入矩阵H0= [v1，…，v|N|]；GCN模型涉及G层，通过以下方式为N 中的节点生成第l层的矩阵：

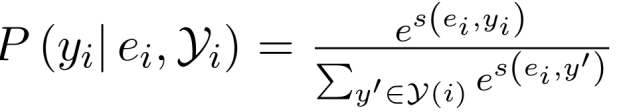
IMG_256

Wl是第l层的权重矩阵；GCN模型的输出是HGH= [h1，…，h|N|]，用作共指预测中节点ni的更抽象的表示向量。

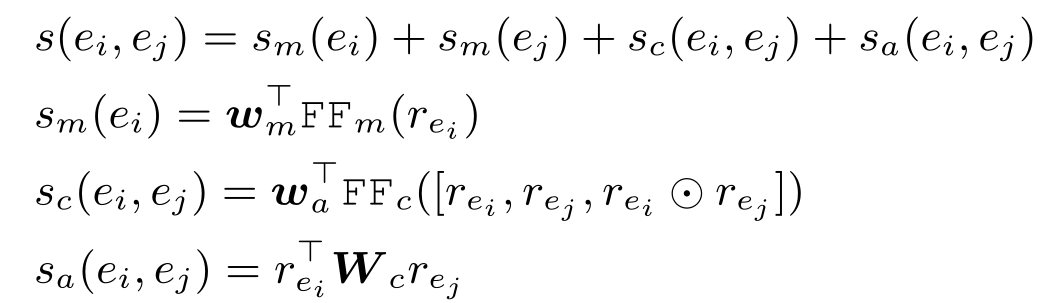
* 1. **端到端共指消解（End-to-end Coreference Resolution）**

首先为每个事件提及ei分配一个先行词Yi∈{e1，…，ei-1，ε），ei到yi中的先行词ej的链接代表了ei和ej之间的共指关系，虚拟先行词表示ei与之前提到的任何事件都不相关；

然后，通过计算预测事件提及ei的共指先行词：

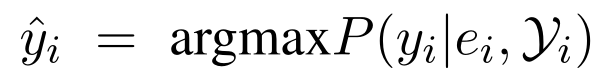


其中s(ei，ej)是一个得分函数，用于确定D中ei和ej之间的共指可能性。通过利用ei和ej的GCN诱导表示向量rei和rej，获得了ei和ej的得分函数s(ei，ej):

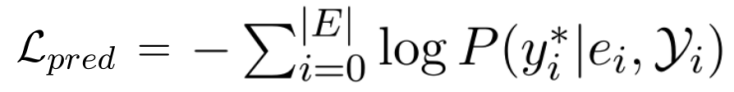


其中，FFm和FFc是两层的前馈神经网络，wm和wa是可学习的向量，Wc是权重矩阵。

最后，ei先行词预测为：



对于训练，作者使用负对数似然作为端到端框架中的损失函数:



* 1. **聚类一致性正则化（Cluster Consistency Regularizaon）**

为了进一步改进ECR的表示学习，作者正则化ECR中事件提及的诱导表示向量，以加强ECR中黄金事件提及聚类和预测事件提及聚类之间的一致性/相似性。

让T = {T1，T2，…，T|T|}和P = {P1，P2，…，P|P|}分别是E中事件提及的黄金集和预测集，对于T或P中的每个聚类C，通过平均事件提及成员的表示向量来为其计算质心向量rC：



{rT1，rT2，…，rT|T|}和{rP1，rP2，… ，rP|P|}分别为T和P的质心向量。使用以下正则化术语来表示聚类一致性:

* **聚类内一致性（Intra-cluster Consistency）**

这种约束关注到每个聚类的内部信息，描述了每个事件提及在它的黄金聚类和预测聚类中的结构。对于每个事件提及ei，我们希望它到对应的黄金聚类T和预测聚类P的质心向量的距离是相似的。

计算ei的表示向量rei到质心向量rT’i和rP’i 的距离：

IMG_256 IMG_256

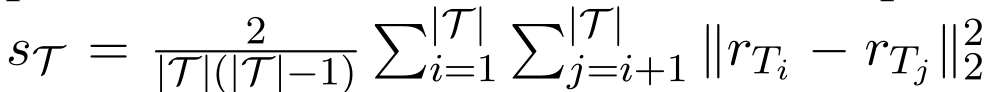
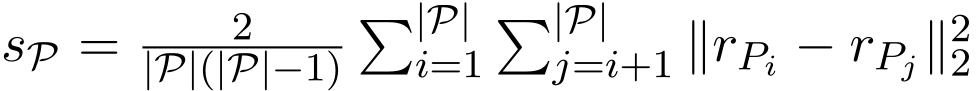
随后，在所有事件提及中汇总黄金聚类和预测聚类的两个距离之间的Linner差，并将其添加到总体损失函数中以实现最小化:

IMG_256

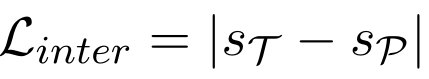
* **聚类间一致性（Inter-cluster Consistency）**

这个约束期望黄金集合T中的簇Ti之间的结构与预测事件簇集合P的结构一致。作者通过聚类的质心向量之间的成对距离的平均值来编码集合中聚类的结构。

黄金聚类T和预测聚类P的聚类间结构分数：

黄金聚类T和预测聚类P的结构得分之间的差异被包括在总损失函数中，以实现最小化:

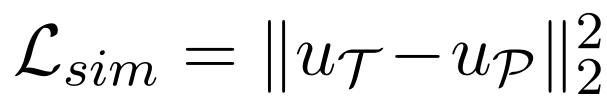


* **集合间相似性（Inter-set Similarity）**

该约束旨在直接促进T中的黄金聚类和p中的预测聚类之间的相似性。对于黄金聚类T和预测聚类P，首先通过平均它们的成员聚类的质心向量来分别获得整体质心向量uT和uP

IMG_256 IMG_256

然后，将Lsim集成到总损耗中，以实现最小化



最后，ECR模型的总损失函数是：

IMG_256

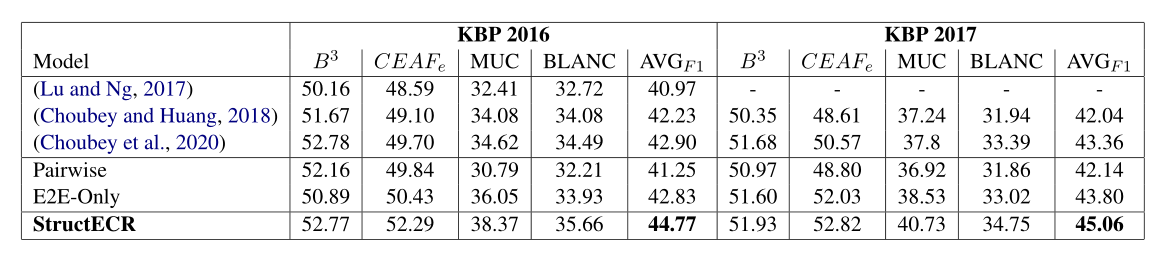
1. **实验及评估**

**部分实验设置**

* **数据集**：
  + 训练：KBP 2015 dataset
  + 评估：KBP 2016 and KBP 2017 datasets
* **评估指标：**
  + MUC、B3、CEAF-e、BLANC、AVGF1
* **参数设置：**
  + G = 2(GCN model)、αinner = 0.1 、αinter = 0.1、αsim = 0.1

**表现评估**

* **对比**
  + ECR和事件检测之间的联合模型(Lu and Ng，2017)；
  + 整数线性规划方法(Choubey and Huang，2018)；
  + 话语结构剖析模型(Choubey et al，2020)；
  + **E2E-Only：**实现了端到端解决模型，文档中的所有事件提及都在单个流程中同时解决，不包括用于表示学习的带有的文档结构组件，聚类一致性正则化也不包括在该模型中；
  + **Pairwise**：没有应用文档结构和正则化术语，Pairwise不是同时解决文档中的事件提及，而是分别预测每对事件提及的共指。



* 消融研究

模型结构中的两个主要组成部分涉及文档结构和聚类一致性正则化。作者进行消融研究，以揭示此类组件对整个模型的贡献。

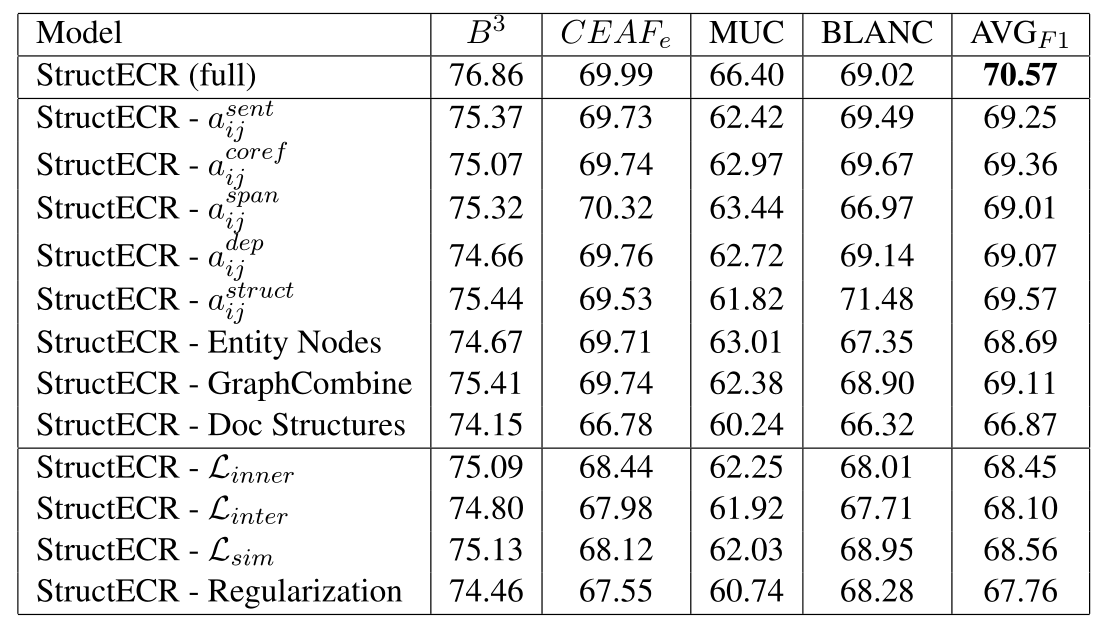
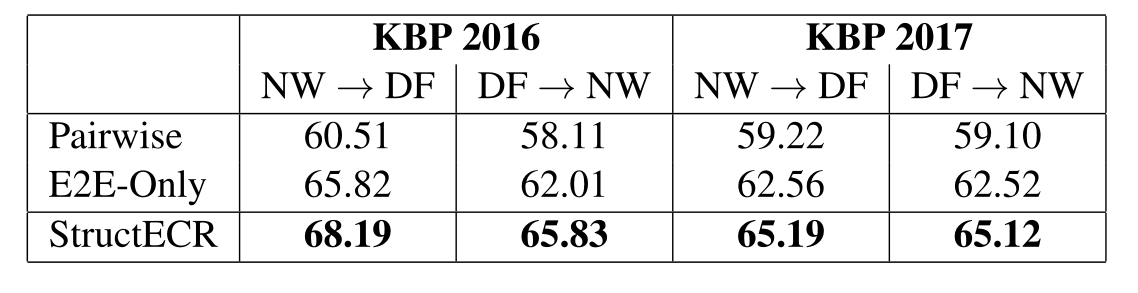


表2显示了模型在KBP 2015年数据集开发数据上的表现。可以看出，从S**tructECR**中删除任何组件都会显著损害性能，表明了S**tructECR**中设计的文档结构和聚类一致性正则化的好处。

* **跨域评估**

目标：在一个域(源域)上训练模型，并在另一个域(目标域)上评估它们。

作者利用KBP 2016和KBP 2017数据集进行实验。KBP 2016年为85篇新闻报道和84篇论坛文档(即两个领域/流派)标注了ECR数据，而KBP 2017年为83篇新闻报道和84篇论坛文档的ECR提供了标注数据。因此，对于每个数据集，作者考虑两种设置，其中一个域(即新闻专线或论坛)中的文档用于源域，而另一个域中的文档用于目标域数据。



从表中可以清楚地看出，在源域和目标域的不同数据集和设置上，**StrucECR**明显且实质上优于基线模型，从而证实了**StrucECR**对于ECR的域泛化优势。

1. **总结**

作者提出了一个基于深度学习的端到端事件提及共指消解框架，主要有两个创新：

* 引入文档结构来显式捕获文档中的相关对象及其交互，以帮助表示学习。
* 提出了几种正则化技术来利用文档中人类提供的和机器生成的事件提及聚类之间的一致性