基于问答匹配的群聊语料角色识别

# 引言

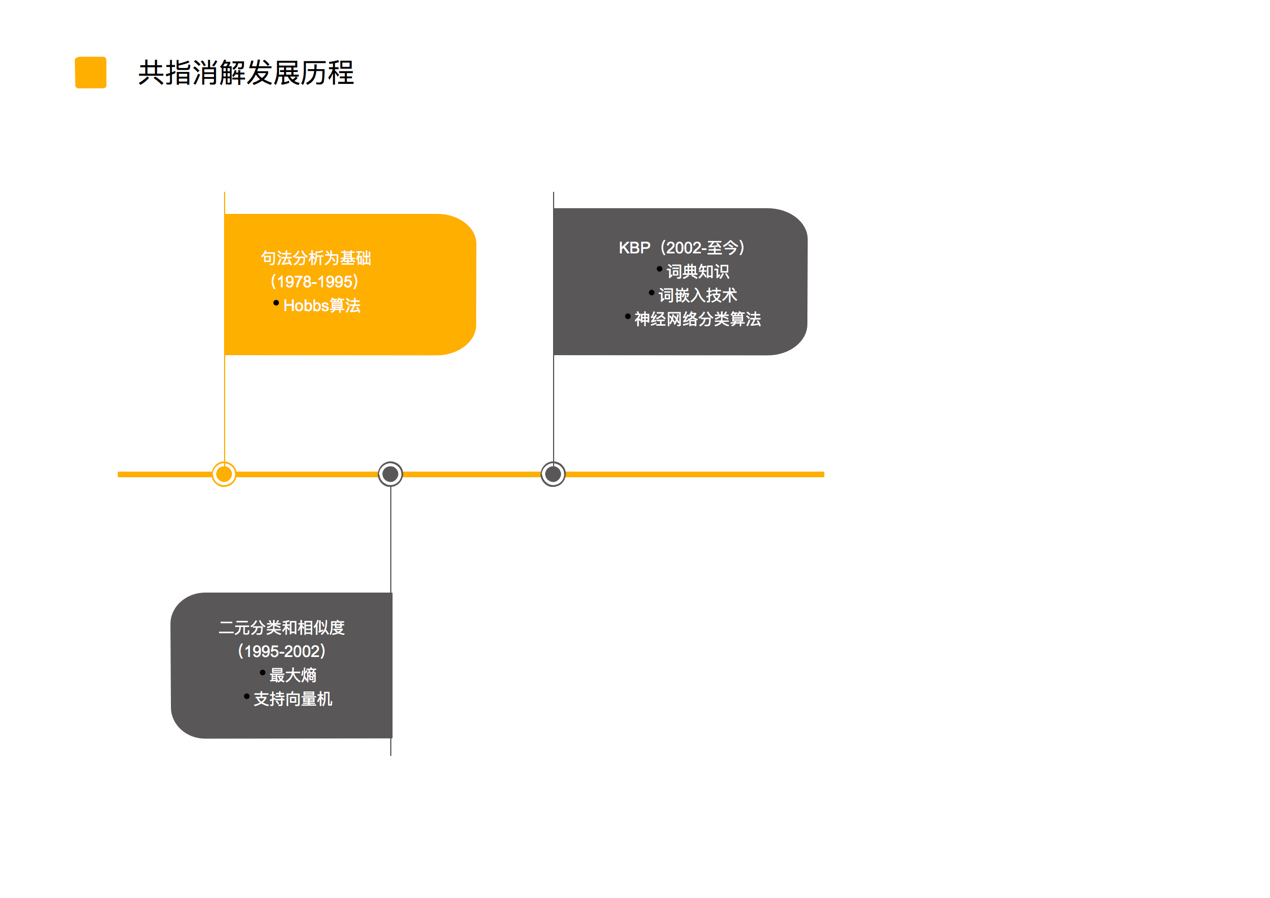
1. 角色识别的目的：角色识别旨在通过多人对话将mention（例如她，妈妈）链接到其指代的全局实体，而全局实体又指向了mention指代的特定角色。
2. 角色识别的必要性：由于多人对话的性质，几个发言者轮流完成一个上下文，因此角色识别是适应高端NLP任务（例如摘要，问题回答，机器翻译等）的关键步骤。
3. 角色识别的难点：对话为口语形式，文本中存在俚语、语法错误等，给角色识别带来更多挑战。（1）普通共指消解是在一篇文本或一个句子中找出名词对应的实体，而角色识别是在多人对话文本中找出mentions对应的角色，mentions与其对应的角色通常不在同一个发言文本中，这就是交叉文档上的共指消解。（2）传统的实体连接可以利用实体知识库中的知识，但角色识别没有关于mentions的先验知识。因此在用向量表达实体时会有一定困难。
4. 相关研究：SemEval2018-Task4的任务为角色识别，网址为：

<https://competitions.codalab.org/competitions/17310#learn_the_details> 。

1. 角色识别应用：（1）问答系统是一种特殊的对话系统，参与对话的人数一般为2人，其中不免也会有一些被提及的人或物，因此识别这些人或物所对应的真实实体对于机器理解文本进而给出合理的回答有重要作用。（2）通过角色识别可以让机器了解关于独立个体的更精细的信息，因此角色识别系统为对话的推断理解提供了帮助。（3）智能助理也需要根据用户的输入信息给出相应的响应，因此系统也需要明确用户输入信息中的mentions所指代的真实实体，而目前的大多数智能助理并没有角色识别功能，只能根据字面义做出相应动作。
2. 群聊语料上的角色识别：指在微信或QQ群聊语料上识别出“@昵称”所指代的特定发言者。与SemEval2018-Task4在剧本语料上的角色识别不同，群聊语料上很少出现能与发言者具有共指关系的实体（比如，发言者的人名，称谓等）。

# 共指消解与角色识别

## 共指消解的历程



## 共指消解与角色识别的关系

照应语（mentions）的定义：文本中所有人名、代词等名词性实体。

* 共指（coreference）的定义：共指指的是两个（或多个）mentions指向真实世界中的同一实体。
* **共指消解（coreference-resolution）**的定义：共指消解就是将指向同一实体的不同mentions连接到同一实体上。
* 实体连接（entity linking）的定义：实体连接可以简单理解为一个分类问题，即给定两个实体和两个实体共同出现的句子文本，预测两个实体之间的关系。
* **角色识别（character identification）**的定义：在一段多人对话中找出被提及的人（人名、人称代词等）所对应角色（说话人）的过程就是角色识别。
* **角色识别与实体连接的关系**：实体连接的目的是预测两个实体之间的某种关系，角色识别的目的是判别代词与其指代的角色之间是否具有指代关系，也可以表述为给定代词预测其指代的某一种或某几种（复数代词）角色。因此，可以说角色识别任务是实体连接任务的子任务。
* **角色识别与共指消解的关系：**两者的共同点是：它们都会对提及的人（mentions）进行分组并标注出每组对应的真实实体。区别在于：（1）角色识别中的共指消解系统需要预测每个mentions组对应的角色（2）角色识别中的共指消解系统处理的是多人参与的对话文本，它是在多个文本之间进行共指消解的。可以认为共指消解是角色识别的一个预处理步骤，利用共指消解找出联系紧密的角色-mentions组。

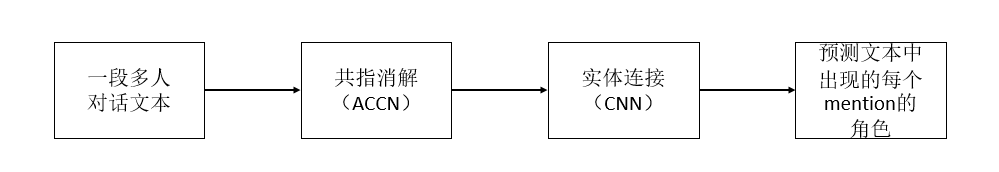
# Emory角色识别系统

## 语料构建



## 系统结构

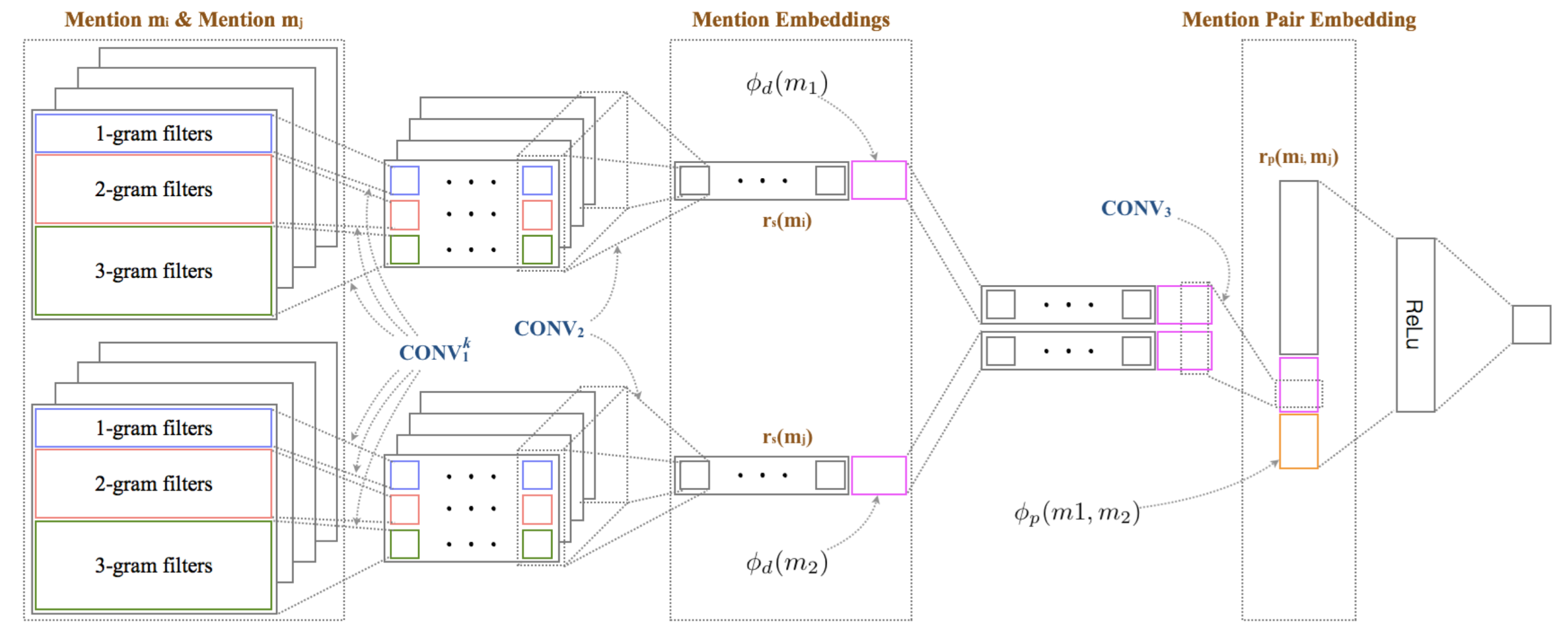
### 整体结构



角色识别系统由两个子系统构成。第1个子系统为共指消解系统。共指消解系统的核心为凝聚卷积神经网络（Agglomerative Convlution Neural Network, ACCN），ACCN结合预训练词向量和语言学特征对mentions、mentions-pair进行嵌入表达，完成共指消解任务。第2个子系统为实体连接子系统。该系统借助共指消解系统从mentions簇中抽取丰富的特征实现角色识别，因此角色识别的效果强烈依赖共指消解的质量。

### 共指消解模型（ACCN模型）

**1. ACCN模型**



**第1部分 计算mention嵌入**

其中，为mention特征模板，抽取mention上下文特征；

为离散特征模板，抽取mention内部特征和说话者特征；

为1号卷积层，对各离散特征向量进行卷积操作；

为2号卷积层，对1号卷积层输出进行卷积操作，||表示concatenate；

2号卷积层的输出作为mention的嵌入向量。

**第2部分 计算mention-pair嵌入**

表示两个mention的嵌入表示的堆叠；

表示3号卷积层，对堆叠后的mention嵌入进行卷积操作；

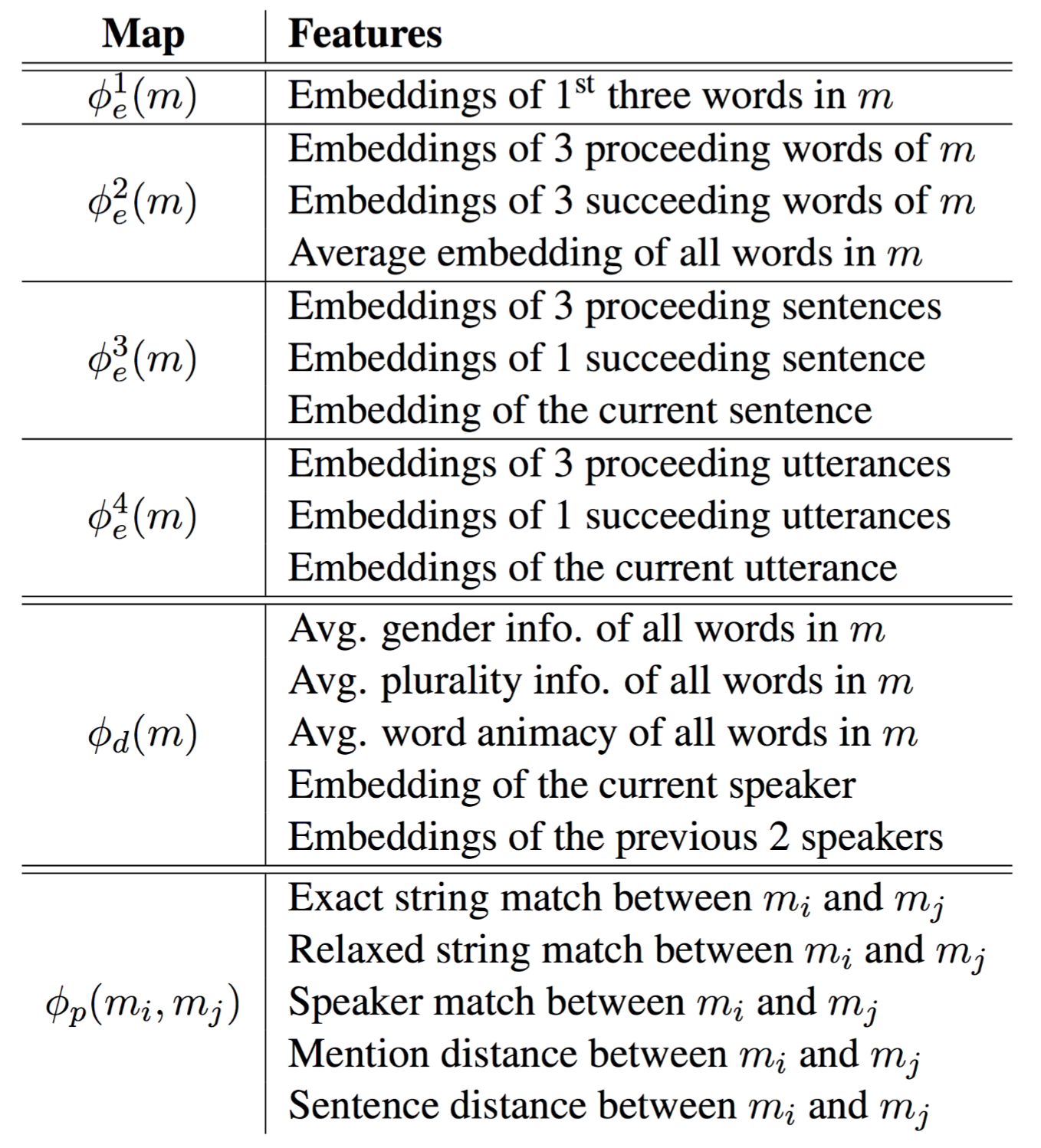
为mention-pair的特征模板，抽取两个mention字面值匹配度，说话者匹配度，距离，所在句子距离；||表示concatenate；

将3号卷积层输出与mention-pair特征向量拼接后作为mention-pair的特征向量。

**第3部分 预测构成共指关系可能性并构建coreference resolution cluster**

Single-Pass思想构建coreference resolution cluster。

**特征模板**



**2. 评测**

**（1）B3**

给定一个mention记为mi，系统给出的结果中包含mi的共指链集合为Smi，标准中包含mi的共指链为Gmi，则mi的准确率和召回率分别为

整体准确率为对所有mention准确率的算术平均值，整体召回率为对所有mention召回率的算术平均值。

**（2）Ceaf**

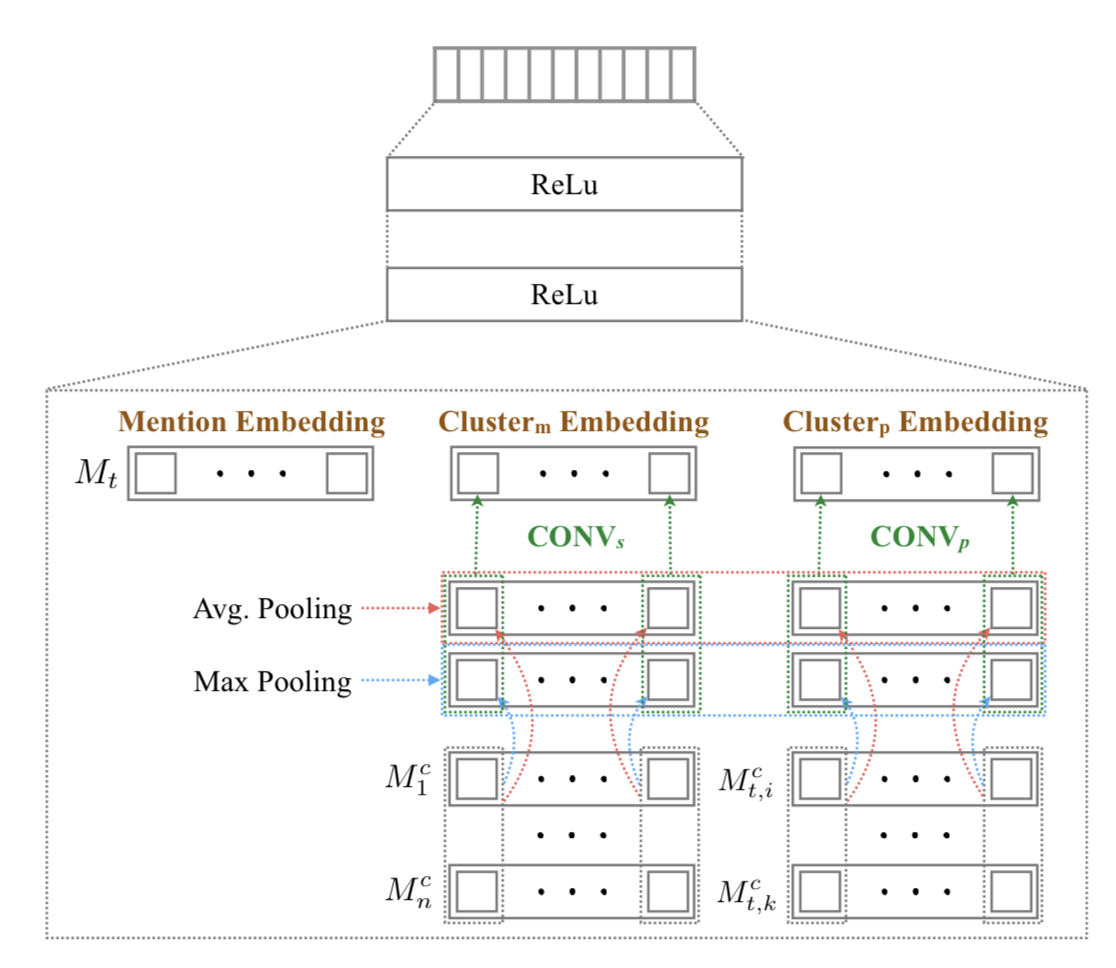
设实体Si表示一个系统给出的实体，Gj表示标准的一个实体，定义

其中表示引用实体Si的mention集合与引用实体Gj的mention集合的交集大小。则系统准确率和召回率为

### 实体连接模型（CNN模型）

**1. CNN模型**

语料经过共指消解系统输出后，得到共指链（簇），得到每个mention的嵌入向量，在实体关系预测过程中，会再次利用CNN神经网络得到每个簇的嵌入向量，结合mention向量送入输出层得到一个mention对应各种角色标记的概率。如下图所示：



图中最上边的两个隐藏层构成一个多类别分类器，下边的部分是构建分类器输入的过程。显然对于一个mention，该模型的下边部分会生成3个向量分别是mention向量，包含mention的簇向量Rs，mention与簇中每个mention构成的mention-pair向量Rs，其中Rp和Rs向量需要各自经过1个卷积和2个池化层得到。

**2. 评测**

（1）用神经网络模型自带的metric度量方法计算准确率。

（2）作者编写的针对每个角色预测的准确率、召回率、F值的评测方法。

（3）作者编写的针对每个角色预测的宏平均准确率、宏平均召回率、宏平均F值的评测方法。

## 中文角色识别系统

### Emory原版系统转换为中文系统

* **语料的修改**

（1）借助翻译API将语料中tokenize后的英文单词序列翻译为中文单词序列。

（2）借助翻译API将语料中transcript后的英文句子翻译为中文。

（3）去掉依存句法标注：将dependency\_tags赋值为null，将dependency\_heads赋值为null。

（4）去掉tokens\_with\_note标注。

* **词典资源的修改**

（1）将英文词向量模型文件替换为中文词向量模型文件。

（2）去掉加载的animate词典，inanimate词典，gender词典，自定义词典。

### Emory原版系统效果

英文的实体连接子系统的效果并没有像论文中所叙述的那样那么理想，说明利用mention向量+mentions簇向量经过卷积层得到的向量来预测mention角色的方法并不理想。

### 中文系统效果

转换为中文的共指消解系统Ceafe法计算的准确率和Blanc法计算的准确率下降了50%，具体原因需要分析。

# 4. 基于问答匹配的群聊语料角色识别

## 疑问句识别

## 基于bert的问答匹配