# Open Domain Event Extraction Using Neural Latent Variable Models Note

# 使用神经隐变量模型进行开领域事件提取笔记

## 摘要

该文考虑的是开放域事件提取，即从新闻集中提取无约束类型的事件的任务。作者构建了一种新的神经隐变量模型，该模型可扩展到非常大的语料库。收集并手动注释数据集，并设计特定于任务的评估指标。

## 以前的工作

从新闻文本中提取事件已引起过很多研究关注。该任务通常由两个子任务组成，即概要归纳（schema induction），用于提取为给定事件类型指定参数槽（argument slot）的事件模板（template）和事件提取，用于从一条新闻中识别带有固定槽（slot）的事件。（一个事件模板（template）与许多槽（slot）有关。比如一个事件由主谓宾组成，则不同事件模板会给主谓宾三个槽填上不同的东西。这个任务反过来就是模板填充。）

以前的建模总是基于给定的事件类型。而金融与安全等领域通常难以完整的给出事件。而且无论如何大多数情况任何机器学习任务我们总有走向无监督的动力。（不过本文的一大假设是事件的报道总是冗余的，如果关注度不高的事件只出现了很少的几次，是很有可能无法提取出来的）

以前的模式归纳工作总是基于人工选取的特征，而本文选取了神经网络中的隐变量(ELMO)进行更好的表征以及神经变分推断策略。这实现了一种新的图模型。这个模型 采用来自一个全局参数化正态分布 的每一个 事件类型隐向量，以及同一实体的文本冗余特性 进行建模。

采用的NPMI来度量槽的一致性。（无监督的工作一定文章中出现度量结果好坏的办法。）

MUC4（1992）是基于槽提取事件的经典方法，关心4种事件（纵火，攻击，轰炸和绑架）以及4个固定槽的组合（犯罪者，工具，目标和受害者）。输入一个新闻输出一个事件，而ODEE则是输入一个新闻集输出一堆事件，并且槽是不固定的。

最早的开创性的工作其实是从基于模式与事件链入手来实现模板提取。而对MUC4最多的策略是采用概率生成模型，

## 挑战

在开放域下事件不再和固定的槽（数量以及种类）有关。这些槽的分布也是不同的。99

在不同新闻中出现的同一实体的情况应该被考虑进来。

We compare our work with Nguyen et al. (2015), the state-of-the-art model on MUC 4 representing each entity as a triple containing a head word, a list of attribute relation features and a list of predicate relation features. Features in the model are discrete and extracted from dependency parse trees. The model structure is identical to our *ODEE-F* except for the features.

惊了，居然他的实验里面有这样一句话。。看来以后不能直接跳过实验。。。：

我们将我们的工作与Nguyen等人进行了比较。(2015)，最先进的muc 4模型将每个实体表示为一个包含头字、属性关系特性列表和pr列表的三元组。编辑关系特征。模型中的特征是离散的，并从依赖关系分析树中提取。除特征外，模型结构与我们的ODEE-F相同。

也做了采用其他特征在odee上。

———— 一言概括：ELMO牛逼。。不过这实验是真的难做。因为它的slot是不固定的，而之前的任务想套用在这上面。。最后只有靠贪心槽映射。整个实验，一定需要许多工具人才能把这个做好，不然一个人明显太消耗时间了。

注意他的event是一个向量建模，所以他又把event向量拿来聚类了。

话说我一直想看下open

Case study：

一个槽在三个事件里出现

事件丢失了槽。（我们对事件的推导只出于与同一谓词相关的足够的多的实体。）

Features in the model are discrete and extracted from dependency parse trees. The model structure is identical to our *ODEE-F* except for the features.

基于实体消歧的生成事件模式归纳

一个槽跳到了另一个事件。（解析树错误传播）

蒸馏学习。