NLP 问题之命名实体识别研究综述

摘　要

自然语言处理是目前人工智能方面的热点问题，其中包含了机器翻译、关系抽取、问答分析等诸多方面的内容。命名实体识别（Named Entity Recognition, NER）则是这些任务的基础工具，为这些任务提供基础服务。

一个文本中有许多具有特定意义的实体，比如人名、地名、组织机构名等专有名词和有意义的时间等，命名实体识别的目标就是根据不同领域或者通用领域的需求识别出不同类型的实体，通常这个过程包括两个步骤：命名实体的边界识别和命名实体的类型分类。

关键字：自然语言处理；命名实体；边界识别；类型分类

Abstract

Natural language processing is a hot topic in artificial intelligence, including machine translation, relationship extraction, question and answer analysis, etc. Named Entity Recognition (NER) is the underlying tool for these tasks, providing the underlying services for these tasks.

There are many entities with specific meanings in a text, such as names of people, names of place, names of organization and meaningful time, etc. The goal of named entity recognition is to identify different types of entities according to the requirements of different fields or common fields. Generally, this process consists of two steps: boundary identification of named entities and type classification of named entities.

Keywords: NLP; named entity; boundary identification; type classification

第一章　引　言

命名实体识别是 NLP 里的一项基本任务，顾名思义，就是指从一个文本中找出命名性指标，通用的指标包括人名、地名、组织机构名这三类。当然，比如在金融、新闻、医疗等特殊领域，也会定义领域内的各种专有实体类型。例如：“小明在北京大学的燕园里看了中国男篮的一场比赛”，这句话中就包括了人名（PER）——“小明”，组织机构名（ORG）——“北京大学”、“中国男篮”，地名（LOC）——“燕园”。如下图 1.1 所示：

小 明 在 北 京 大 学 的 燕 园 里 看 了 中 国 男 篮 的 一 场 比 赛

PER　　ORG　　LOC　　ORG

图 1.1 中文命名实体识别

相对于英文，中文命名实体没有明显的形式标志，还存在分词的干扰，导致中文命名实体识别难度也高于英文。现如今我们使用的实体检测与识别的途径主要有两种，一种是先进行实体检测，再去对已经检测的实体进行识别，另一种是将实体与识别的对象结合到一个模型里，同时得到字符的位置进行标记和类别标记。

第二章　中文命名实体识别的困难

引言部分也已经提到，汉语作为一种象形文字，相比于英文等拼音文字来说，针对中文命名实体识别任务来说往往更复杂，其难点主要存在于以下几个方面：

（1）　由于汉语的语法规则导致中文实体结构比较复杂，并且有些类型的实体词的长度没有限制，不同的实体有不同的结构，比如组织机构名存在大量嵌套：铁岭小马餐饮管理有限公司，其实是一个嵌套的、完整的组织机构名，而不应该识别成地名，人名，组织机构名；人名中也有较长的少数民族人名或翻译过来的外国人名，由于构词没有统一的规范，所以这类命名实体识别的召回率相对偏低。

（2）　中文文本不像英语文本一样，天然的有空格作为词语的界限标志，而且在汉语中“词”的概念很模糊，不同语境有不同的“词”，上下文不同时，同一词可能就会产生不同的实体类型，中文也没有英文中的大小写等形态指示。

（3）　别名、缩略词的问题；中文里广泛存在简化表达现象，如“东大”、“医大一院”等名词。

（4）　在不同的文化、领域、背景下，命名实体的外延存在差异，即较难实现一个很

好的通用领域的命名实体识别系统，通常仅限于某个专业领域内。

第三章　命名实体识别的发展

就像目前很多的深度学习模型和任务一样，比如机器翻译等，从历史上看都大致分为三个主要阶段：基于规则的阶段，基于统计的阶段，基于神经网络的阶段。现在回过头来看，三个阶段的发展是一脉相承的，在不同的历史条件和当时的硬件资源限制下，计算机界的专家们在其特定时代，提出了在当时比较适用的方法。

命名实体识别从历史上看也存在三种解决方法：基于规则的方法、基于统计的方法、基于神经网络的方法。

3.1 基于规则的方法

这种方法需要各个领域的专家们总结命名实体的规则，并由计算机专家手工编写这些规则，将文本与规则进行匹配来识别出命名实体。例如，对于中文来说，“说”、“老师”等词语可作为人名的下文，“大学”、“医院”等词语可作为组织机构名的结尾，还可以利用到词性、句法信息。

但是，由于文本数量庞大，语言规则复杂，专家们对于某些规则的不一致看法，规则与规则之间还可能存在冲突等，导致在构建规则时，其过程费时费力、可移植性不好。此外，还要不断更新规则库，所以这种方法代价太大。

3.2 基于统计的方法

统计机器学习的方法将命名实体识别视作序列标注任务，利用大规模语料来学习出标注模型，从而对句子的各个位置进行标注，主要方法包括：隐马尔科夫模型、最大熵、支持向量机、条件随机场（Conditional Random Fields, CRF）等。

在上面提到的四种学习方法中，最大熵模型具有较好的通用性，结构紧凑，但是它的训练时间复杂度很高，有时甚至会导致训练代价难以承受，另外由于需要明确的归一化计算，导致开销比较大。一般说来，最大熵和支持向量机在正确率上要比隐马尔科夫模型高一些，但是隐马尔科夫模型在训练和识别时的速度要快一些，主要是由于在利用维特比算法求解命名实体类别序列的效率较高。所以，隐马尔科夫模型更适用于一些对实时性有要求以及像信息检索这样需要处理大量文本的应用，如短文本命名实体识别。

在统计时代，解决命名实体识别的流行做法是特征模板 + CRF 的方案。特征模板通常是人工定义的一些二值特征函数，试图挖掘命名实体内部以及上下文的构成特点。对于句子中的给定位置来说，特征的位置是一个窗口，即上下文位置。而且，不同的特征模板之间可

以进行组合来形成一个新的特征模板。

CRF 是是一种判别式概率模型，是随机场的一种，常用于标注或分析序列资料，如自然语言文字或是生物序列。简单是说在 NER 中应用是，给定一系列的特征去预测每个词的标签，如下图 3.1 所示。

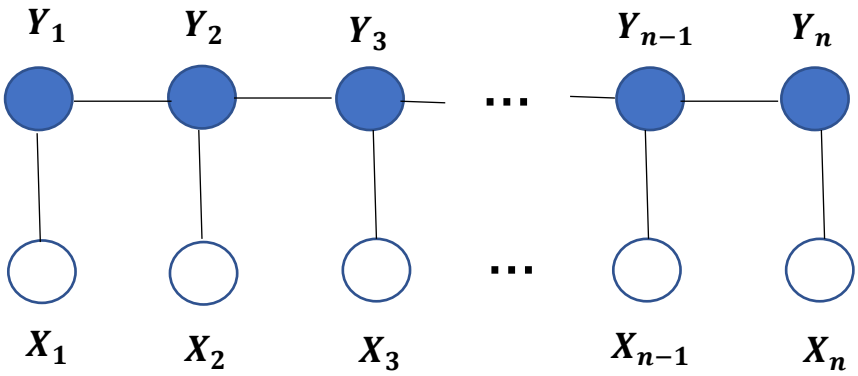

图 3.1 CRF 中的某一序列

X 我们可以看作成一句话的每个单词对应的特征，Y 可以看作成单词对应的标签。这里的标签就是对应场景下的人名、地名等等。CRF 的优点在于其为一个位置进行标注的过程中可以利用到此前已经标注的信息，利用 Viterbi 解码来得到最优序列。对句子中的各个位置提取特征时，满足条件的特征取值为 1，不满足条件的特征取值为 0；然后把特征喂给 CRF，训练阶段建模标签的转移，进而在预测阶段为测试句子的各个位置做标注。

基于统计的方法对特征选取的要求较高，需要从文本中选择对该项任务有影响的各种特征，并将这些特征加入到特征向量中。依据特定命名实体识别所面临的主要困难和所表现出的特性，考虑选择能有效反映该类实体特性的特征集合。主要做法是通过对训练语料所包含的语言信息进行统计和分析，从训练语料中挖掘出特征。有关特征可以分为具体的单词特征、上下文特征、词典及词性特征、停用词特征、核心词特征以及语义特征等。

基于统计的方法对语料库的依赖也比较大，而可以用来建设和评估命名实体识别系统的大规模通用语料库又比较少。但是，在当时的确是一种比基于规则更好的方法。

3.3 基于神经网络的方法

近几年来，随着硬件资源的飞速进步，以及词的分布式表示（word embedding）概念的出现，使得前人们曾提出过的神经网络一跃成为解决诸多问题的主流方法，其中就包括 NLP 任务。

这类方法对于序列标注任务（如 CWS、POS、NER）的处理方式是类似的，将词（token）从离散的独热（one-hot）表示映射到低维空间中成为稠密的词嵌入（embedding），随后将

句子的 embedding 序列输入到 RNN 中，用神经网络自动提取特征，Softmax 来预测每个 token 的标签。这种方法使得模型的训练成为一个端到端的整体过程，不依赖特征工程，是一种数据驱动的方法；但网络变种多、对参数设置依赖大，模型可解释性差。此外，这种方法的一个缺点是对每个 token 打标签的过程中是独立的分类，不能直接利用上文已经预测的标签（只能靠隐状态传递上文信息），进而导致预测出的标签序列可能是非法的，例如标签 B-PER 后面是不可能紧跟着 I-LOC 的，但 Softmax 不会利用到这个信息。

为了解决上述的独立分类问题，学术界又提出了用长短期记忆模型-条件随机场（LSTM-CRF）做序列标注。也就是说，在 LSTM 层后加入 CRF 层来做句子级别的标签预测，使得标注过程不在是对各个 token 独立分类。

目前经常使用 Bakeoff-3 评测中所采用的 BIO 标注集，即 B-PER、I-PER 代表人名首字、人名非首字，B-LOC、I-LOC 代表地名首字、地名非首字，B-ORG、I-ORG 代表组织机构名首字、组织机构名非首字，O 代表该字不属于命名实体的一部分。如下图 3.3 所示：

小	明	在	北	京	大	学	的	燕	园
B-PER	I-PER	O	B-ORG	I-ORG	I-ORG	I-ORG	O	B-LOC	I-LOC

图 3.3 BIO 标注

当然也可以采用更复杂的 BIOSE 标注集。

经过多次的尝试，目前解决命名实体识别流行的做法是利用 BiLSTM + CRF 框架，如下图 3.4 所示：

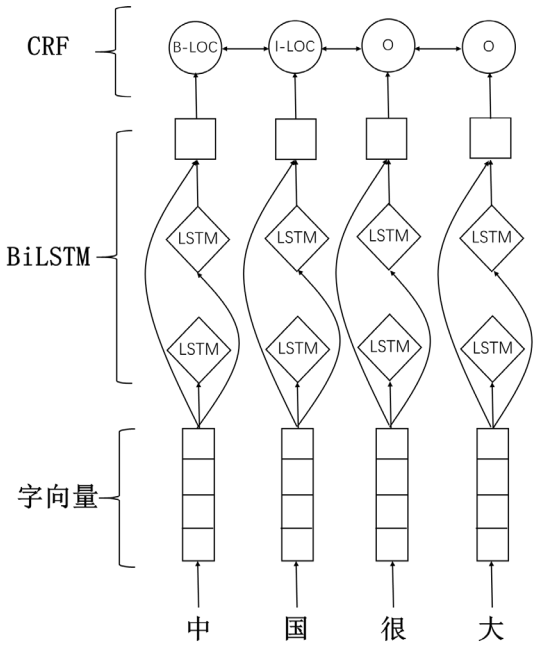

图 3.4 命名实体识别模型 BiLSTM-CRF

第四章　命名实体识别的评价

作为一项任务，命名实体识别当然需要一个或几个指标来作为对效果的衡量标准。目前主要根据两个评价指标衡量命名实体识别系统的性能：召回率和准确率。如图 4.1 所示：

		Actual class: 1	Actual class: 0
Predict class	1	True Positive	False Positive
	0	False Negative	Ture Negative

图 4.1 真实与预测分类

根据上面的表格，我们可以得到召回率和准确率的公式如下：

$$Precision = \frac{True\ Positives}{\#Predicted\ Positives} = \frac{TP}{TP + FP}$$

$$Recall = \frac{True\ positives}{\#Actual\ Positives} = \frac{TP}{TP + FN}$$

由上面的公式，我们可以肯定，拥有高准确率或者高召回率的模型是一个好的模型，所以我们会尽量去提高模型的准确率和召回率。但是在实际过程中发现，精确率与召回率这两个指标通常在大规模数据集中时候相互制约的，并不能同时得到提高，这就需要综合考虑，最常见的方法就是 F-Measure，它是精确率和召回率加权调和平均，其计算公式如下：

$$F_\alpha = \frac{(\alpha^2 + 1) * Precision * Recall}{\alpha^2 * (Precision + Recall)}$$

其中α是参数用来表示精确率与召回率关注程度的不同，Precision 和 Recall 分别通过前面的公式就算得出。当$\alpha = 1$时，就是命名实体识别中常用的评价指标，即：

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

第五章　总结与期望

由于本人能力有限，所以文中没有涉及代码部分的实现，当然还有很多不足之处。本文只介绍了中文命名实体识别中的监督学习方法，一些半监督、无监督的方法还未介绍。而且随着大规模预训练模型的提出，如 Bert，命名实体识别的性质出现很大的提升，本文也未做介绍。本文只是详细介绍了单模型框架下的各部分的情况，而对集成学习（ensemble）方法与模型选择策略未做描述。今后，随着个人能力的提升，我会不断丰富中文命名实体识别研究综述。

在本学期的课程中，我也学到了很多，通过肖老师的讲解，我对 NLP 入门也有了一些了解，在课上通过和其他人交流学习，也努力的去配置实验环境，并且成功调通了学长给的程序代码，对于初级的我来说还是很有积极作用的。最后一堂课，听了学长学姐们的分享，也让我了解到了生活中的一些经验和一些应有的做事态度：不要啥都说不会…等等。

最后，感谢朱老师、肖老师，以及实验室的学长学姐们的帮助，虽然我现在是个专业小白，但是我也会尽自己的努力，争取向这些优秀的人学习，不只是学习上。我要学习的东西还很多，未来继续努力吧！再次谢谢朱老师、肖老师的辛苦付出！