命名实体识别综述

作者：王占坤 班级：07111706 学号：1120170124

一 、命名实体识别简介

1. 命名实体识别
   1. 命名实体

命名实体是一个能够清晰地从与其具有相似属性的集合中进行定义区分的词语或者短语。 (Sharnagat, 2014)

* 1. 命名实体识别概念

命名实体识别（英语：Named Entity Recognition, 简称NER），是指识别文本中的具体特定意义的实体。命名实体识别旨在从文本中提取实体并将其分类到预定义的分类集合当中。NER在NLP中有大量的应用，如信息抽取、问答系统、机器翻译等。 (维基百科)

1. 研究内容
   1. 研究对象

命名实体识别的研究主体一般包括三大类和七小类。三大类分别是实体类、时间类和数字类。七小类分别是人名、地名、机构名、日期、货币和百分比。命名实体识别就是识别文本中的这三大类、七小类的实体。命名实体识别的任务可以分解成两个子任务即实体边界（Entity Boundaries）确定和实体类别(Entity Types)划分。

其中，由于时间、货币和百分比等具有较为规范的规律，可以依据正则表达式等进行识别。而人名、地名、机构名等用字较为灵活，识别难度比较大。所以命名实体识别的主要研究领域也就是对人名、地名、机构名等的实体的识别。 (孙镇，王慧临, 2010)

* 1. 实体领域

在实际应用过程中，还需要根据具体的应用场景来确定。不同的应用场景需要对不同的实体识别进行优化。比如，在面向教育领域，经常将餐厅、学校、教师、学生等作为命名实体；面向体育领域，经常会将分数、跑步、篮球等作为命名实体。在这种情况下，一种领域的模型在应用于另一领域时就会出现严重的性能下降。

* 1. 语种

命名实体识别还要依据不同的语种进行特别的调整。例如，由于英文文本在命名实体的识别过程中只需要考虑词本身的特征而不涉及分词问题，实现难度相比中文就比较低。中文中的ER的生成规律、语法结构、应用环境等相较于英语就更加复杂，很难通过简单的设置规则或者简单的神经网络进行特征。中文的命名实体识别任务也因此变得很艰巨。因此不可能使用一种模型应用于所有语种的命名实体识别任务，必须要根据语种特点对模型加以修改和优化。

1. 发展历程

NER最初在第六届信息理解会议（MUC-6）上被提出 (Grishman R, 1996)。此后，研究者对NER进行了深入的研究。依据时间线，我们可以将NER方法大致分为以下四种： (Jing Li, 2018)

① 基于规则的方法。依赖于手工设置的规则，不需要标注的数据

② 无监督学习的方法。依赖于非监督机器学习算法，不需要标注的数据。

③ 基于特征的监督学习。依赖于监督学习算法以及优良的特征表示。

④ 深度学习。一种端到端的策略，可以自动探测所需的分类以及特征表示。

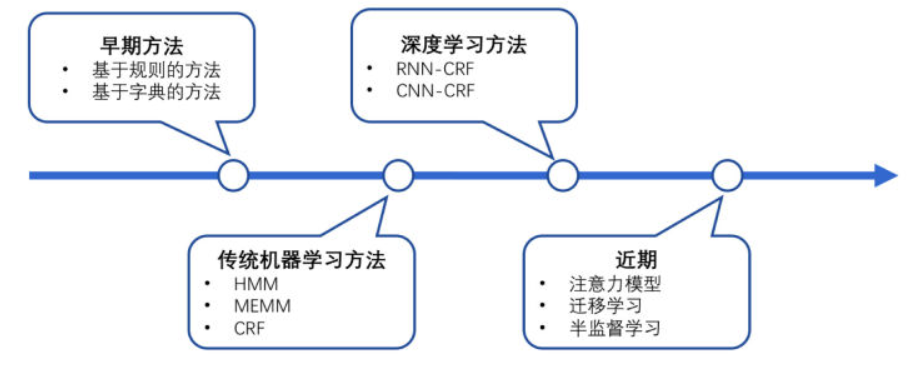


Figure 来自<https://zhuanlan.zhihu.com/p/43061858>

1. 国内外研究

命名实体识别在自然语言处理中占据很重要的位置，也是不可逾越的学术问题。自命名实体识别任务提出以来，国内外都对此进行了深入的研究并形成了一定的成果，产生了很多专门的dataset以及toolkit。

* 1. 研究机构

国内外关于命名实体识别的主要研究机构和相关工作如下： (白宁超, 2019)

(1) 国外研究机构主要研究英语等语言的实体识别，代表机构包括斯坦福研究所人工智能中心、因特尔研究中心、微软研究院、雅虎研究中心、日本东京大学等。

(2) 国内主要解决中文命名实体识别任务，代表机构包括中科院计算所、微软中国研究院、哈尔滨工业大学自然语言处理实验室、北京语言大学语言信息处理研究所、北京理工大学自然语言处理研究室和复旦大学自然语言处理研究室等。

4..2 Datasets

目前命名实体识别领域的主要数据集有：MUC-6, MUC-6 PLUS, MUC-7, CONLL03, ACE等。这些标注好的数据集会在NER算法研究过程中帮助研究者验证和训练。这些训练集各有其针对领域。如NYT（The New York Times Annotated Corpus）就收录了1987年1月1日至2007年6月19日纽约时报撰写和发表的文章。该集合包含超过650,000个文章 - 摘要对，这些数据对有助于自动文档摘要的算法的开发和评估。

* 1. Off-the shelf Tools

目前，已经产生了很多NLP工具。如：StanfordCoreNLP, OSU Twitter NLP, NERsuits，精准自然语言解析器（SyntaxNet）,中文自然语言处理工具包（FudanNLP）, Open NLP, 自然语言处理工具包NLTK、CRF++等。不同的NLP工具有其一定的特点和特征。例如，NLTK能够为50多种语料库和词汇资提供易于使用的方法，并且提供了一套用于分类、标记、词干化、解析和语义推理的文本处理库，和用于工业级NLP库的包装器。 (nltk)

二、评价方式

在南洋理工大学Jing Li等人的论文中，将评价方式分为了两类，第一类是完全评价匹配，第二类是宽松评价。

1. 完全评价匹配

所谓完全评价匹配就是指实体边界和实体类型都符合要求。完全匹配评价的参数主要有三个Precision, Recall和F-score。这些参数是通过计算TP,FP和FN得来的。

1. TP：被NER识别为实体且正确识别
2. FP: 被NER识别为实体但识别错误
3. FN：被标注数据标注的实体但是没有被NER识别出来

①

②

③ (Sang E F, 2003)

其中公式三当取β=1时转化为F1score

④

所谓Precision就是正确识别的比率。Recall就是正确识别的占应该正确识别的比例。F1值就是Precision和Recall的均值，根据需要可以自行调整权重。

2. 宽松评价

根据Jing Li 等人的论文描述：ACE提出了一个更复杂的评估程序。它解决了部分匹配和错误类型等问题，并考虑了命名实体的子类型。然而，这是有问题的，因为只有当参数固定时，最终得分才具有可比性。复杂的评估方法不直观，难以进行错误分析。因此，复杂的评估方法在最近的NER研究中没有被广泛使用。

ACE (Doddington G R, 2004)在论文中提出的评价公式如下。分数被定义为整个系统输出实体的综合。

④

⑤

三、命名实体识别方法

在命名实体识别简介中的第三部分发展历程中，简单提到了在NER历史上主要的发展历程。下面个将对这几种方法进行简述。这里的方法分类采用国内普遍使用的分类，即：基于规则、基于统计学习、基于深度学习的命名实体识别方法。

1. 基于规则

Farmakiotou D (Farmakiotou D, 2000)等人在发表于2000年的论文中对基于规则的命名实体识别技术进行了介绍。基于规则的命名实体依赖于手工设置规则。规则主要是基于语义学、领域专有名词等进行设置。通常包含在NER系统中的词汇资源是词典，以列表字典的的形式保存。现有的NER模型系统可以分为以下类型：

1. 基于手工制作的语法和地名录，也即是字典。典型的有LaSIE, LaSIE II, FASTUS
2. 利用机器学习技术自动获取NER词汇资源
3. 结合前两种方法的系统，如爱丁堡大学的LTG系统

基于规则的方法命名实体识别的方法的缺点是相当明显的。首先是需要人工花费大量时间对语言学进行研究，才能构建出合适的规则。其次，这种方法仅适用于小数据集。当数据集逐渐增大时，采用设置规则和总结字典的弊端就会显现出来。数据集越大，无论是词义还是实体类型都变得异常复杂，以至于很难设计出较好的规则。

但是，另一方面，基于规则的NLP技术为后续NLP技术的发展提供了大量的数据字典，方便了后续技术的研究和发展。

就目前而言，研究者仍然需要通过手工设置规则、制作数据字典等方式提高NER的准确率。特别是在应对诸如生物学、物理学、医学等含有大量专有词汇的学科时，数据字典可以较好地提高准确度。

1. 基于统计学习

鉴于基于规则的方法有耗时长、难度高等弊端，人们逐渐将目光对准统计学领域，希望通过统计学的方法设计NER技术。

基于统计机器学习的方法本质上就是一种多分类的方法。刘浏 (刘浏，王东波, 2018)指出，在进行命名实体识别研究中，首先要给定命名实体的多种分类，然后再使用模型对数据中的实体进行分类。

基于统计机器学习的命名实体方法具有可移植性好、语言依赖性小、处理速度快等优点，主要是通过序列标注实现的。 (白宁超, 2019)

白 (白宁超, 2019)总结了命名实体识别问题想序列标注问题的转换的主要处理步骤：

1. 统计学习策略：选择合适的机器学习方法。常用的有：隐马尔可夫模型（HMM）、条件随机场模型（CRF）、最大熵模型（MEM）等。
2. 特征选择：特征的选择直接影响命名实体识别的好坏。
3. 模型训练：通过对训练集的训练优化算法模型。
4. 模型评测：训练出来的算法模型，通过测试集进行测试。

隐马尔可夫是一个用来描述一个含有未知参数的马可夫过程的统计模型，是一个关于时序的概率模型。隐马尔可夫模型已经得到广泛应用，并在应用中取得了很好地效果。但是，隋臣 (隋臣, 2017)指出隐马尔可夫模型具有不能有效利用上下文信息的缺点。条件随机场CRF也是假设状态满足马尔科夫性，但是相比于HMM，他可以引入更多特征，能够记录包括词语本身和词语所在的上下文的特征。

目前采用更多的是将基于统计的机器学习方法与基于规则的命名实体识别方法结合起来。这样既能够通过统计的方式降低人工设计规则的成本，另一方面通过手工设置规则的方式可以提高准确率、召回率和F值。

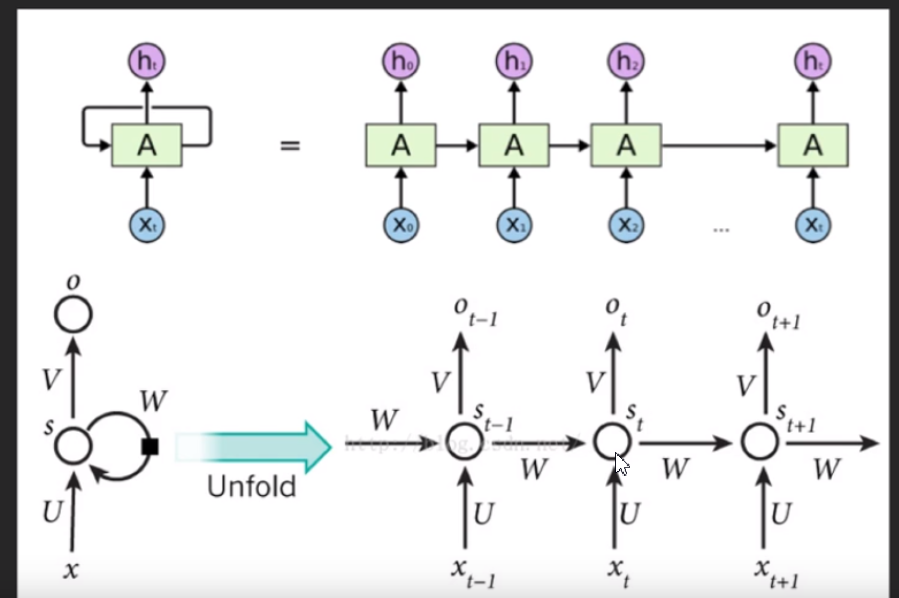
但是基于统计的缺点就在于需要大量标注好的数据。

1. 基于深度学习

深度学习是基于深层神经网络Deep Neural Network, DNN) 的学习方法的别称 (刘树杰，董力，张家俊, 2015)。深度学习是近年来机器学习领域发展最为迅速的。随着对深度学习的深入研究，基于深度学习的NER模型逐渐成为NER模型研究的重点。相比于传统的NER方法，深度学习方法可以自动进行隐层特征的识别，可以自动对数据进行特征识别和抽象。深度学习是一种端到端的学习过程，他隐藏了学习的过程中的参数调整，使得可以专注于对神经网络本身进行研究和优化，无需关心在数据训练过程中参数的变化，也即是无需关心对特征的提取与识别。深度学习的优势就在于可以使用非监督学习或半监督学习和分层特征提取高效演算法来替代手工取得特征。

在基于深度学习的NER中， (Jing Li, 2018) RNN是最广泛使用的上下文解码器之一，CRF常用语标签解码器；BiLSTM-CRF则是基于深度学习NER方法的最常见的架构。

3.1 RNN



2来自https://www.youtube.com/watchv=7uWGxBH5m40&list=PLua227btV7cTc26SD9kZfGRrQfRzy2K1o&index=7

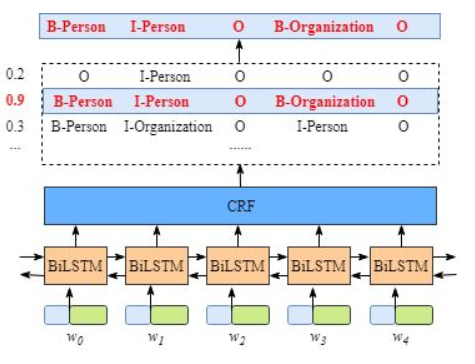
如上图所示是RNN的结构模型。RNN主要用来处理序列数据。在NER中，我们经常需要通过上下文来准确预测判断一个词的词义、词性。RNN可以做到记忆当前词之前的词。

RNN相比于传统神经网络，其hidden layer的同层神经节点是相互连接的，即一个序列当前的输出与前面的输出也相关。当前，基于RNN的LSTM模型已在NLP领域获得巨大成功。

3.2 BiLSTM-CRF

循环神经网络RNN能够很好地处理序列数据。在NLP中，我们往往需要知道一句话中前面的词以预测后面的词的词义，也即是神经网络需要具有记忆性。由于RNN会产生后面结点对前面结点的感知力下降的问题，人们又设计了LSTM。LSTM可以选择性的对某些信息进行记忆和遗忘。BiLSTM在LSTM的基础上有对未来信息进行了记忆。BiLSTM由三层组成即是：look-up层，双向LSTM层，CRF层。

BiLSTM-CRF的输入时词向量，输出则是每个单词的标签，在命名实体识别中也即是对实体进行分类。首先是BiLSTM层，它的输入时词向量表示和对应词句，通过look-up层得到对应的词向量输入；它的输出是对应单词的分数，这些得分将作为CRF层的输入。在CRF层，我们可以加入一些约束条件进行改进来保证结果是正确的



3来自<https://zhuanlan.zhihu.com/p/44042528>

四、挑战与未来发展

NER自提出以来，一直是NLP领域的一个重要研究方向和不可逾越的学术问题。从MUC到ACE在到CONLL一系列重要会议既提出并逐渐明确了NER的研究范围，也提出了大量经典的评测方法和研究工具，也提出了很多具有重要意义的问题。 (刘浏，王东波, 2018)从NER研究方法来看，经历了从基于规则到基于统计机器学习再到深度学习的阶段，NER的发展也是机器学习发展的一个缩影。研究方法逐渐从严重依赖人工标记、设置规则再到部分依赖数据字典的监督机器学习再到完全由深度学习神经网络自动提取词向量特征、词性特征进而对命名实体进行识别，实现了从人工到机器、从被动到主动、从小数据到大数据的转变。2018年BERT的提出和今年CMU全新XLNET等预训练语言模型的出现，使得我们现在有机会在不需要进行复杂特征工程的情况下对NER进行进一步的研究和发掘NER的未来潜在的方向。

1. 挑战

无论是有监督的NER学习模型还是基于深度学习的模型都需要大量标注好的文本数据进行训练和测试。然而，数据标注仍然耗时且昂贵。特别是在一些专业领域，需要数据字典辅助训练以提高模型效果。

除了标注数据语料的缺失，还有命名模糊、语义混淆等的问题。由于不同训练集之间的标注的不一致，也导致在训练模型本身迁移性比较差。

除此之外还有不同语言的由于语言学上的障碍导致很难设计出一个通用的跨语言NER模型。

1. 方向

(Jing Li, 2018) Li等人指出了未来NER发展的几个方向：细粒度NER和实体边界确定，实体链接，DL-based NER on Informal Text with Auxiliary Resource，基于深度学习的NER的拓展，NER的深度转移学习，易使用的基于深度学习的NER模型。

五、总结

本文简要回顾了NER的发展历史，并对其中的几种NER方法进行了概要。NER既是历史上NLP发展的重要领域，未来随着深度学习的发展，以及NLP的需求的增多，NER仍将是NLP的重要领域并将原来发展的基础上提出更多更新的基于深度学习的通用模型。未来具有更强通用性的系统方法可能成为研究重点。

# 参考文献

Doddington G RA, Przybocki M A, et al.Mitchell. (2004). The Automatic Content Extraction (ACE) Program-Tasks.

Farmakiotou DV, Koutsias J, et alKarkaletsis. (2000). Rule-based named entity recognition for Greek financial texts.

Grishman RBSundheim. (1996). Message Understand ing Conference-6: A Brief History[C].

Jacob DevlinChang, Kenton Lee, Kristina ToutanovaMing-Wei. (2018). BERT: Pre-training of Deep Bidirectional Transformers for.

Jing LiSun, Jianglei Han, and Chenliang LiAixin. (2018). A Survey on deep learning for Named entity Recognition.

nltk. (无日期). Natural Language Toolkit. 检索来源: nltk: http://www.nltk.org/

Sang E FMeulder F.De. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition.

Sharnagat. (2014). Named entity recognition: A literature survey.

白宁超. (2019年2月). 命名实体技术. 检索来源: https://bainingchao.github.io/2019/02/13/%E5%91%BD%E5%90%8D%E5%AE%9E%E4%BD%93%E8%AF%86%E5%88%AB%E6%8A%80%E6%9C%AF/

刘浏，王东波. (2018). 命名实体识别研究综述.

刘树杰，董力，张家俊. (2015). 深度学习在自然语言处理中的应用. 中国计算机学会通讯.

隋臣. (2017). 基于深度学习的中文命名实体识别研究.

孙镇，王慧临. (2010年6月26日). 命名实体识别研究进展综述. 现代图书情报技术, 页 42-47.

维基百科. (无日期). 命名实体是被. 检索来源: 维基百科: https://zh.wikipedia.org/wiki/%E5%91%BD%E5%90%8D%E5%AE%9E%E4%BD%93%E8%AF%86%E5%88%AB

祖木然提古丽·库尔班，艾山·吾买尔. (2019). 中文命名实体识别模型对比分析.