RC23590 (C0504-006) April 22, 2005
Computer Science

IBM Research Report

A Combination Training Framework for Domain-Specific Nominal Entity Recognition

Hong Lei Guo, Zhi Li Guo
IBM Research Division
China Research Laboratory
HaoHai Building, No. 7, 5th Street
ShangDi, Beijing 100085
China

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

基于模型组合训练机制的特定领域名词性实体识别

郭宏蕾 郭志立

IBM 中国研究中心

Email: {guohl, guozhili}@cn.ibm.com

摘要 本文提出了一个模型组合训练机制，用于建立特定领域名词性实体识别模型。该组合训练机制在特定领域名词性实体识别模型的训练中，采用基于文本片段的语料库自动构建机制，从 Web 的搜索结果中挖掘所需要的领域数据， 并充分利用已有的一般领域名词性实体识别模型、标注语料库及自动新建的小规模的特定领域名词性实体标注语料库，极大地降低了训练成本，为特定领域名词性实体识别模型的建立提供了一个简单易用的训练方法。

A Combination Training Framework for Domain-specific Nominal Entity Recognition

Guo Hong Lei and Guo Zhi Li

IBM China Research Lab, Bei Jing

Email: {guohl, guozhili}@cn.ibm.com

Abstract In this paper, we present a combination training framework for building a domain-specific nominal entity recognition model. To reduce the huge cost in domain-specific corpus collection and tagging, this combination training framework leverages the existing nominal entity tagged corpus and nominal entity recognition model built in general domain. Meanwhile, a web-based automatic corpus construction mechanism is applied to collect the domain-specific data from the search results on the web. Experimental results show that this combination training framework can significantly reduce the training cost in building a domain-specific nominal entity recognition model. It provides an ease-to-use way for building a domain-specific nominal entity recognition model with less time and efforts.

1 引言

名词性实体识别主要关注识别表述人、地点、组织的名词短语，这类短语允许嵌套命名实体。名词性实体识别是实体识别和跟踪应用中的一个基本任务，已成为自动内容抽取评测(ACE)中的一个基本评测点[1]。

机器学习方法因其易于训练和调整的特点在信息抽取、命名实体和名词性实体识别研究领域倍受青睐。近年来举行的各种命名实体识别系统评测（如ACE和CoNLL）显示多种有监督的机器学习方法已成功地用于命名实体识别并取得令人满意的效果[2]。尽管基于机器学习的实体识别系统易于调整到新的领域，但有监督的学习方法在训练识别模型时对标注语料库的依赖性非常大，通

常需要利用大规模的标注语料库来克服数据稀疏问题以获得较好的性能。然而，由于建立大规模手工标注语料库通常非常昂贵， 目前世界上可用的标注语料库还非常少，特定领域的标注语料库则更少，特定领域标注语料库已成为建立基于有监督机器学习算法的特定领域名词性实体识别系统的一个主要瓶颈。

每个领域的名词性实体都有一些特定的内部语言学特征和外部语境，当基于机器学习方法的名词性实体识别系统用于不同于原训练领域的新领域时，其性能通常会有所下降。因此，在实际应用中，人们希望能找到一个有效的训练调整方法，只需花费较少的时间和精力就可以将已有的一般领域名词性实体识别系统迅速调整到一个新的特定领域，既能满足特定领域的识别需求又不降低识别性能。

本文提出了一个构建特定领域名词性实体识别模型的组合训练机制。该组合训练机制充分利用已有的一般领域名词性实体识别模型和自动构建的特定领域语料库，以较少的时间和精力训练特定领域名词性实体识别模型，极大地降低了训练成本。在特定领域名词性实体标注语料库构建中， 我们采用了一个基于文本片段的语料库自动收集方法， 从Web上收集相关领域数据，用于建立一个小型特定领域名词性实体标注语料库。实验结果表明，该组合训练机制只需花费较少的时间和精力就能对已有的一般领域中文名词性实体识别系统进行快速调整，以满足特定领域需求。

本文组织如下:第二节提出了一个用于构建特定领域名词性实体识别模型的组合训练模式。第三节描述了自动构建特定领域名词性实体标注语料库的方法。第四节给出了几种减少自动标注错误所带来的噪音的方法，并分析了相关的实验结果。第五节给出了结论。

2 面向特定领域名词性实体识别的组合训练机制

当基于机器学习的名词性实体识别模型用于新的领域时，其性能通常会有所下降。一些研究人员已开始关注为特定领域建立自适应的基于机器学习的实体识别系统[3][4]。我们提出一个组合训练机制，只需花费较少的时间和精力就可构建一个特定领域名词性实体识别模型（见图 1）。该机制包括两个部分：1）特定领域语料库的自动收集和标注； 2）模型组合训练。

在训练特定领域名词性实体识别模型时，通常需要建立一个大规模的特定领域名词性实体标注语料库， 这需要投入大量的时间和精力，大大增加了训练成本。为了减少语料库的构建成本， 我们采用了一个语料库自动收集和标注方法（见第 3 节），从 Web 上搜集相关的领域数据，用于建立特定领域标注语料库。训练语料的质量和数量是影响识别模型的重要因素，由于自动标注的语料库的质量通常比手工标注的语料库低，为了保证模型学习的有效性，在模型训练中，我们还集成了已有的一般领域名词性实体识别模型和手工标注语料库， 这既可以从自动标注的特定领域语料库中挖掘与特定领域名词性实体相关的内部语言学特征和外部语境信息，又能从已有的一般领域标注语料库中捕获名词性实体所共有的基本内部特征和外部语境信息，减少自动标注错误带来的数据噪音。这样，我们只需花费较少的时间和精力就可建立一个大规模的训练集，用于学习特定领域名词性实体识别模型。

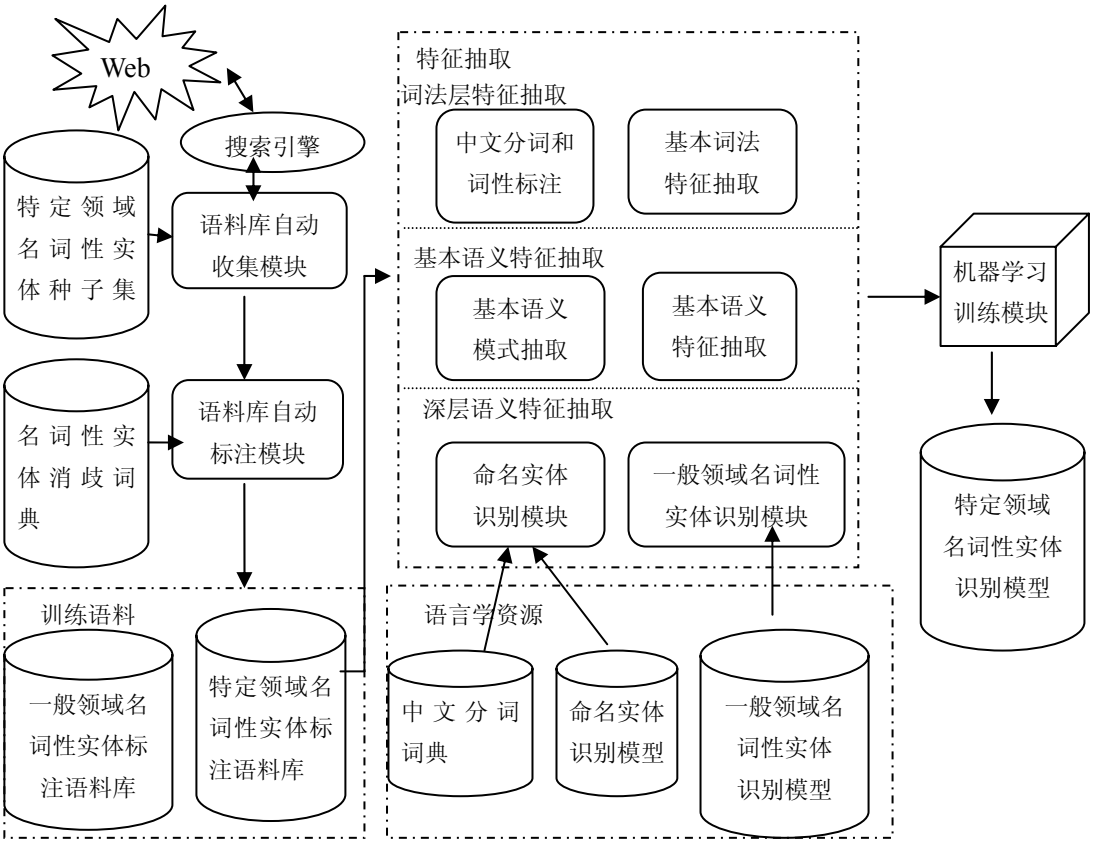

图 1 特定领域名词性实体识别模型的组合训练机制

在模型化特定领域名词性实体任务时，我们采用一个模型组合训练机制，充分利用已有的一般领域名词性实体识别模型所提供的名词性实体识别结果和其他语言学特征。其他语言学特征主要包括中文词法特征（如汉字、中文分词单元、 词性）、基本语义特征（如时间、日期等特定语义结构、基本语义类型）、深层语义特征（如命名实体、名词性实体核心片断）。该组合训练机制，只需要较少的时间和精力即可对已有的一般领域中文名词性实体识别模型进行调整，以满足特定领域的名词性实体识别需求。

3. 特定领域名词性实体标注语料库的自动构建

建立大规模的特定领域名词性实体标注语料库通常需要更多的领域知识，是一项非常艰巨的任务。我们采用了一个基于文本片段的语料库自动构建方法，以减少特定领域名词性实体识别模型的训练成本。

Web 可看作一个无限的语言资源库，含有各种出现在不同语境下的多种名词性实体。由于从各网站上收集所有相关的文档全文需要花费很多精力，我们建立了基于文本片段的语料库自动构建机制，利用一个特定领域名词性实体种子集和搜索引擎从 Web 上搜索相关的领域数据，然后从返回的 Web 搜索结果中收集含有名词性实体种子的文本片断， 自动标注文本片断中所含的中文名词性实体示例。这样，我们无需花费太多精力和时间就能收集和标注更多含有特定领域名词性示例的文本片断， 有助于减少特定领域名词性实体识别中的数据噪音，克服数据稀疏问题。

本节以金融领域为例，阐述了名词性实体标注语料库的自动构建方法。为了克服自动标注歧义，特定领域标注语料库的自动构建包括两步：1）基于文本片段的特定领域语料库的自动收集；2）特定领域名词性实体的自动标注。

3.1 基于文本片断的特定语料库自动收集

从 Web 上收集含有名词性实体种子的全文文档通常需要花费很多的时间和精力， 因此，我们主要收集至少含有一个特定领域名词性实体示例的句子。预定义的特定领域名词性实体种子集主要选自金融领域辞典[5], 共包含 227 个表述人的名词性实体和 334 个表述组织的名词性实体。在语料库自动收集中，将这些名词性实体种子作为查询关键字提交给搜索引擎（如 http://www.google.com），每个查询能返回 10~1000 个搜索结果， 从这些返回的搜索结果中抽取相关文本片断。为了减小数据噪音，我们还对这些文本片断进行分句并进行冗余数据的过滤处理，移走重复和无意义的句子、所有元数据信息（如目标文档的 URL、搜索引擎所提供的注释和其他基本元数据）和冗余数据（如重复的句子和搜索结果）。 我们利用这些经过求精过滤后的文本片断构建了一个特定领域名词性实体的生语料库。

3.2 特定领域名词性实体的自动标注

名词性实体识别模型中所用的特征包括名词性实体的内部特征及其外部语境特征。构建特定名词性实体标注语料库目的就是从真实文本中挖掘特定领域名词性实体的特定语言学信息。为了减少标注成本，我们对所收集的文本片断中的特定领域名词性实体进行了自动标注，其过程如下。

1) 建立名词性实体消歧词典

为了减少自动标注中的错误，我们建立了一个名词性实体消歧词典。该词典由一般领域名词性实体词表、命名实体词表和特定领域名词性实体种子集组成，总计约 55，245 个实体词条，其中，一般领域名词性实体词表由已有的手工标注的一般领域名词性标注语料库中的全部名词性实体及其内部片段组成，命名实体词表由该一般领域名词性标注语料库中的全部命名实体组成。此处所用的一般领域名词性实体标注语料库的文章均选自 2001 年-2002 年的“北京青年报“的新闻报道， 共有 2M 汉字。

2) 特定领域名词性实体的自动标注

在特定领域名词性实体的自动标注中, 我们使用名词性实体消歧词典和基于分词单元的最大精确匹配法对样例中的候选实体词条进行自动标注。如果样例所含的某个分词单元序列能与消歧词典中的候选实体词条精确匹配，则对该分词单元序列进行相应的标注。 基于分词单元的最大精确匹配能过滤掉基于字符的最大匹配方法中经常出现的子串匹配错误。

根据上述步骤， 我们构建了一个特定领域名词性实体标注语料库(约 335924 个汉字)。

我们称至少含有一个特定领域名词性实体示例的样例为一个正样例。我们抽取出上述特定领域标注语料库中的所有正样例，构建了一个特定领域正样例数据集（约 312807 汉字）， 共含有自动标注的名词性实体 24142 条，该正样例数据集将用于训练特定领域名词性实体识别模型。

尽管自动标注的语料库比手工标注的语料库质量低，但我们的特定领域名词性实体标注语料库的构建过程是完全自动的， 所花费的时间和精力都较少，并且随着 Web 上的数据资源的不断增长，无需花费太多成本就可对语料库加以扩充。

4. 采用组合训练方法降低自动标注错误的数据噪音影响

4.1 实验数据

在特定领域名词性实体识别模型的训练中，我们使用了三个训练数据集：Gen_m、, Dom_auto、和 Dom_m (见表 1)。Gen_m 是手工自动标注语料库；Dom_auto 是采用第 3 节中所述的语料库自动收集和标注方法建立的特定领域名词性实体自动标注语料库；Dom_m 则是在自动收集的特定领域生语料库上手工标注的名词性实体标注语料库。测试数据分为一般领域测试数据集和特定领域测试数据集两部分（见表 1），测试数据集中的所有名词性实体均由手工标注，以确保测试数据的正确性和测试结果的可靠性。一般领域测试数据集均选自 2001-2002 年的北京青年报的新闻报道， 特定领域测试数据集则由 15 个银行法律法规组成。目前，我们的名词性实体识别模型主要用于识别表述人的名词性实体（PER）和表述组织的名词性实体(ORG)。

在特定领域名词性实体标注语料库构建中，我们使用了基于文本片段的语料库自动收集方法，这可获取更多信息丰富的训练样例和特定领域上下文信息，有助于训练更具有自适应能力的特定领域名词性实体识别模型。为了衡量标注语料库中的名词性实体分布状况， 我们定义了“名词性实体密度“，名词性实体密度定义为每一千个汉字中含有的名词性实体个数。从表 1 中，我们可看出，基于文本片段的小规模特定领域语料库 Dom_m（0.278M 汉字）的名词性实体密度远远高于大规模的一般领域语料库 Gen_m（2M 汉字）, 这表明基于文本片段的语料库自动收集方法可以在一定程度上缓解特定领域名词性实体的数据稀疏问题。当然，自动标注语料库 Dom_auto 的标注质量目前还低于手工标注语料库 Dom_m，自动标注错误主要来源于自动标注中名词性实体的边界定位错误。在训练特定领域名词性实体识别模型时，为了减少自动标注错误的噪音，我们采用了多种组合训练方法。

表 1 训练数据集和测试数据集

数据集 (标注方法)	语料库大小	名词性实体数目	名词性实体密度
Dom_auto (自动标注)	0.278M	22,951	82.58
Dom_m (手工标注)	0.278M	15,635	56.24
Gen_m (手工标注)	2.0M	44,819	22.41
特定领域测试数据集	0.093M	3,691	39.68
一般领域测试数据集	1.35M	34,475	25.54

4.2 基本实验

实验中所用的一般领域名词性实体识别模型是从 Gen-m 语料库中训练建立的，其在一般领域中已取得令人满意的性能（见表 2）。 然而，当将该模型直接应用于金融领域时，其性能明显下降（见表 3）。显然，该一般领域名词性实体识别模型在特定领域的自适应能力是非常有限的。

表 2　一般领域名词性实体识别模型的性能

名词性实体类型	准确率 (%)	召回率 (%)	F 值 (%)
PER	88.20	87.12	87.66
ORG	83.77	77.88	80.72
All	86.88	84.26	85.55

表 3 一般领域名词性实体识别模型在金融领域的性能

名词性实体类型	准确率(%)	召回率 (%)	F 值(%)
PER	78.16	63.14	69.85
ORG	87.24	80.53	83.75
All	84.83	75.4	79.87

表 4 基于特定领域语料库 Dom_auto 的特定领域名词性实体识别模型的性能

名词性实体类型	准确率(%)	召回率(%)	F 值 (%)
PER	57.28	71.96	63.79
ORG	60.26	79.86	67.55
All	59.40	75.43	66.46

我们还利用自动标注的特定领域语料库 Dom_auto (0.278M)建立了一个特定领域名词性实体识别模型，其基本性能也较差（见表 4）。显然，自动标注的错误严重影响了该识别模型的准确率。

4.3　利用组合训练方法降低自动标注错误的数据噪音

为了减少自动标注错误带来的数据噪音，尽可能地捕获名词性实体的基本语言学特征和语境信息， 我们在特定领域名词性实体识别模型建立中充分利用了其他各种可用的资源，例如小型手动标注的特定领域标注数据集、已有的高质量的一般领域名词性实体识别模型和一般领域名词性实体标注语料库。实验结果显示这些资源能进一步减少自动标注错误带来的噪音，改进特定领域名词性实体识别模型的性能。

4.3.1 利用一般领域名词性实体识别模型减少自动标注错误的数据噪音

在基于模型组合训练的特定领域名词性识别模型构建中，我们将一般领域名词性实体识别模型的识别结果作为一维特征参与训练，这为特定领域名词性实体识别模型提供了更多名词性实体所共有的基本语言学特征。 在这个模型组合训练实验中，训练数据集由 Gen_m 和 Dom_auto 组成。实验结果（见表 5）显示基于模型组合的新的特定领域名词性实体识别模型的 F-值提高了 1.13 个百分点。由于自动标注的领域数据中含有更多的特定领域名词性实体及其所在的语境信息，每个名词性实体类型的召回率也有所提高。然而，精确率却因为受标注错误引入的噪音的影响而有所下降。

如果在训练中不利用一般领域名词性实体识别模型的输出结果，则所建特定领域识别模型的性能有所下降（见表 6）， 其 F 值和准确率分别下降了 1.11 个百分点和 2.5 个百分点。实验结

果表明一般领域名词性实体识别模型所提供的识别结果为特定领域名词性实体识别提供了重要的线索， 能有效捕获更多名词性实体的基本内部语言学特征和外部语境特征，减少自动标注错误产生的数据噪音。

表 5 基于模型组合训练机制的特定领域名词性实体识别模型的性能

名词性实体类型	准确率(%)	召回率 (%)	F 值 (%)
PER	74.57	81.15	77.72
ORG	83.22	81.37	82.28
All	80.50	81.31	80.90

表 6 未使用模型组合训练机制的特定领域名词性实体识别模型的性能

名词性实体类型	准确率(%)	召回率(%)	F 值(%)
PER	75.07	79.39	77.17
ORG	79.23	82.59	80.88
All	78.00	81.66	79.79

4.3.2 利用小规模手工标注的特定领域名词性实体语料库减少数据噪音

在训练中， 我们还引入一个小规模手工标注的特定领域语料库 Dom_m*(0.1M)，用以减少自动标注错误所产生的数据噪音（见表 7），使训练所得的特定领域名词性实体识别模型的性能显著提高，其 F 值可达到 83.79%。

表 7 基于组合训练机制的特定领域名词性实体识别模型的性能

训练数据集 (数据集大小)	F 值(%) （未组合一般领域名词性实体识别模型）	F 值(%) （组合一般领域名词性实体识别模型）
Gen_m (2M)	79.87	----
Dom_auto (0.278M)	66.46	69.83
Gen_m (2M) + Dom_auto (0.278M)	79.79	80.90
Gen_m (2M)+ Dom_auto* (0.178M) + Dom_m* (0.1M)	82.65	83.79
Dom_m (0.278M)	88.28	88.65

注释：1）Dom_m*(0.1M)是从 Dom_m 中随机选取的 0.1M 领域数据；2）Dom_auto* (0.178M) 是从 Dom_auto 中随机选取的 0.178M 领域数据；

如果用手工标注的特定领域语料库 Dom_m (0.278M)替换自动标注的语料库 Dom_auto (0.278M)， 基于模型组合训练集机制构建的特定领域识别模型的性能(F=88.65%)比一般领域名词性实体识别模型在金融领域的识别性能（F=79.87%）提高了 8.78 个百分点（见表 7），而且对从未出现在训练集中的新名词性实体的识别性能可提高 30.06 个百分点。这些重大的性能改进表明标注语料库的质量是影响名词性实体识别系统质量的一个重要因素。

即使不借助一般领域名词性实体识别模型和一般领域名词性实体标注语料库，仅用手工标

注的特定领域语料库 Dom_m (0.278M)进行训练，所得的特定领域名词性实体识别模型的性能也能显著改进，其 F 值达到 88.28%。这表明我们的语料库自动收集机制能有效地收集到信息丰富、质量较高的领域数据，而且领域数据的覆盖率和平衡性均比较好。

上述试验结果表明：

1）在组合训练机制中，只需要投入较少的时间和精力建立一个小规模的特定领域语料库，这极大地降低了特定领域名词性实体识别模型的训练成本并获得较好的识别准确率。

2）基于文本片段的语料库自动收集和标注机制能有效地收集信息丰富的领域数据，大大减少了构建特定领域标注语料库所需的时间和精力。

3）尽管自动标注语错误对特定领域名词性实体识别模型的训练产生了一些副作用，我们可利用已有的一般领域名词性实体模型、手工标注语料库有效地降低自动标注错误所产生的负面影响。

5. 结论

本文给出了一个模型组合训练机制，用于建立特定领域名词性实体识别模型。该机制在特定领域名词性实体识别模型的训练中， 充分利用了已有的一般领域名词性实体识别模型、标注语料库及自动新建的小规模的特定领域名词性实体标注语料库，这极大地降低了训练成本，使训练成本减少为原来的 50%。

特定领域名词性实体标注语料库是名词性实体识别模型训练中的一个主要瓶颈。我们采用了一个基于文本片段的语料库自动构建方法，从 Web 的搜索结果中挖掘所需要的领域数据，构建特定领域名词性实体标注语料库。实验结果表明，利用小规模的特定领域名词性实体标注语料库和其他已有的一般领域训练数据，只需花费较少的时间和精力就可训练出一个令人满意的特定领域名词性实体识别模型。这种模型组合训练机制简单易用，能快速调整一般领域名词性实体识别系统，使其满足特定领域名词性实体识别任务的需求。

当然，如何用最少的时间和精力建立特定领域名词性实体识别模型仍是摆在我们面前的一个巨大挑战，我们将进一步探索有效的样例选择方法，用于挖掘更多信息丰富的特定领域训练样例。

6. 参考文献

1. Chinese Annotation Guideline for Entity Detection and Tracking (version 4.2.4), 2004
2. Erik F. Tjiong Kim Sang and FienDe Meulder. 2003. *Introduction to the CONLL2003 shared task: Language independent named entity recognition*, Proceedings of CONLL2003.
3. Jianhan Zhu, Victoria Uren, Enrico Motta, ESpotter: Adaptive Named Entity Recognition for Web Browsing, Proc. of Workshop on IT Tools for Knowledge Management Systems, 2005
4. Fei Huang, Alex Waibel, An Adaptive Approach to Named Entity Extraction for Meeting Applications, Proceeding of HLT 2002.
5. 葛玲芝, <<华尔街金融词典>>, 天津大学出版社, 2004.