文章编号：2096-1618(2017)增-0021-04

跨语言无监督依存分析方法研究

胡 华，郭 非
(四川中烟工业有限责任公司成都卷烟厂，四川 成都 610000)

摘要：跨语种转移方法用于依赖关系已经取得了良好的效果。虽然这种方法使用的是依存树库的目标语言，但是，大多数方法都仍在使用大型平行语料库。据报告，并行数据是语言的稀缺资源，新方法不需要并行数据，内嵌语法的学习方法概括一个双语的语法语境词汇，并将这些结合成神经网络分析器。在给基线改进展示的同时，解析器是使用依存树库的数据集的。分析源的重要性是语言文字，并表明了是源语言的组合导致了基合金的改进。

关 键 词：跨语言；无监督学习；依存分析；源语言

0 引言

依存分析是许多自然语言处理系统中的关系抽取、统计机器翻译化、文本分类和在线答疑的一个重要组成部分。基于监督办法的依存分析在众多语言中的应用已经非常成功。然而，对于许多其他语言，依存分析仍旧不可用，并且是非常昂贵的。这激发了无监督学习方法的发展，可以利用未注释的单纯的数据，使用无监督方法来提高精度。

在对于资源少的语言使用无监督工作的依赖性解析中，采用了解析虚化和跨语言的转移方法。在此，解析器是在一个训练好的资源中，直接把一个资源处理的目标语言。这里唯一的要求是，得到语言的 POS 标签时必须使用相同的标记集。这个假设是对相关资源的语言而言的。此外，还有许多高精度 POS 的报告标注的资源语言。在跨语种虚化的方法中已经出现了比无监督更好的方法。并行数据可用于一个跨语种解析器。然而，并行数据可能很难获得的真正的资源。因此，我们建议的方法能够改善跨语种的性能。

解析器是使用一种语言的数据。我们的做法是基于充实的解析器来做句法字嵌入。在句法方面，源语言和目标语言都映射到基于共享的低维空间并且都需要等位基因的数据。前人的研究主要集中在合并字嵌入信息的分析模型，我们提出了一种利用神经网络的方法，其中大多数的研究是通过选择性使用高资源的语言数据来模拟的低资源。因此数据是丰富的，例如，只有很少的并行数据的存在(例如，词典，维基百科)。采用两个阶段来训练我们的分析器：先学习跨语言的句法字嵌入，那么学习的其他参数作用到源语言树形解析模型。当作用到目标语言的时候，发现所有的研究语言所得到的结果是一致的。这项工作的目标是建议一个通用解析器，这样可以解析很少语言。

当存在多个源语言的时候，可以尝试通过选择最好的源语言或由几个源语言组合而成的源语言来提高性能。前人已经提出了一种用于选择给出最佳的源语言和目标语言的装置。为了解决这个问题，引入了两个指标，其中基线作为源语言。还提出了结合所有可用的方法源语言，从而导致大量的证法。

1 非监督跨语种

依存分析对于资源贫乏的语言解析器的建立两种主要的方法：解析和凸起。泽曼提出了虚化的方法。他们建立了一个从树库中解析的源语言。这个解析器可以训练任何标准的监督方式，包括任何词汇特征，然后在资源贫乏的句子语言中应用二解析从。解析虚化是依赖于事实的，并且这部分的词性是具有悬垂关系的。例如，在英文判别圆弧因素中，依赖解析器实现了84.1%的准确度，而版本达到78.9%。建一个使用丹麦语和瑞典语的解析器，这两种语言是紧密相关的。还采用跨语言词簇作为其解析器的虚化功能，从语言集抽取的联合建模。相比之下，投影方法是利用数据项目源语言与目标语言的依赖关系。给定一个源语言的分析树，所产生的靶子通过投影的语言来解析树。然而，方法是依赖于许多启发式的，这将难以适应其他语言。是利用虚化解析和并行数据，用英语虚化分析器作为种子解析器的目标语言，并根据字对齐进行更新。该模式鼓励目标语言语法树，这种树是类似于源语言分析树的。使用并行数据来传输的源解析器受限

收稿日期：2016-07-04

于目标端的比对。对于空对齐,用虚化解析器,而不是源语言词汇化解析器。

综上所述,现有工作一般是创建一个虚化分析器,并且使用并行数据信息来改善它。相反,要提高解析器的虚化,但不使用并行数据或者任何明确抽动资源。

2 虚化改进解析

提出了一种新的方法,这种方法是用来改善一个跨语言解析器无追索权的并行数据。方法没有使用额外的资源。该方法是基于语法字嵌入,一个字是作为句法的低维向量空间。这个想法很简单:我们将要虚化的解析器使用文字嵌入,其中,源语言和目标语言词汇项目在同一空间表示。词中通常在嵌入的同时捕获同步战术和语义信息。但是,对于悬垂解析而言,字嵌入是需要反思的。在接下来的小节中,重新查看一些跨语种字嵌入,并提出句法字嵌入。第 4 节在比较这些字时并入了依赖解析器。

2.1 跨语言文字嵌入

可以代表单词的是在一个包括源语言和目标语言的低维空间。相反编号是用来表示尺寸等于词汇量的大小,字是用小很多的尺寸来表示。缓解数据稀疏的问题是通过编码词汇学习的关系来交涉。有几种方法都力求于交叉并行数据字嵌入。基于这样的理念,表示的杆等位基因的句子应该是紧靠在一起的。他们构成的句子级表示为袋,然后对铰链损失函数进行了优化。虽然这看起来非常适合作为一个单词表示在跨语言解析中的需要,它可能会导致过度语义嵌入,这是反式重要特征,但对于分析用处不大。例如,两个具有不同的表达的词有不同的表示抽动功能。在每种情况下,这种高亮显示的标签是用来强调字的预测的。建立一种使用跨语种字来表示布朗聚类器的变体,这种变体是用来作用于并行数据的。研究表明,对于依存分析,基于简单的布朗簇算法要优于很多字嵌入技术。在文中,比较了几个形成跨语种字的方针。

2.2 句法字嵌入

现在,提出了学习跨语种字嵌入的新方法是着重于语法的。文中的嵌入方法是利用字共现以及建筑上分布的传统技术的相似性,例如,词语的共现是围绕中心词的上下文的。不是相邻的标记,使得捕捉头部存在嵌入和修改关系。他们指出,这一策略比表面嵌入提供了更好的单语依存分析。然而,他们的方法是不适用于我们的低资源设置,因为它需要训练解析树。相反,我们考虑一个简单的表示,即部分讲话的上下文。这仅需要词性标注,而不是完全的解析,同时如果在 POS 的背景下提供语法信息链接,期望能够为信息的依赖关系。算法 1 句法字嵌入:

(1)匹配源和目标统一到企业通用标记集。

(2)为提取序列字源语言和目标语言。

(3)对于每一个 n 元,保持中间字,并其 POS 更换等字样。

(4)从嵌入模型的源语言和目标语言上所产生的字和 POS 序列名单都假设标记集

用于两个相同的 POS 源和目标语言,在算法 1 中,学习两种语言的单词嵌入和每个单词类型附近的同样的语法空间文本。特别是,我们开发预测模型标签的左侧和一个字的权利。从源语言英语和目标语言西班牙语,其中突出的显示的片段体现了周围的每一个重点单词预测。注意对于本实例中,POS 上下文的中文和西班牙语的动词是相同的,有几个方法:

(1)基于 POS 标签的过于粗粒度精确解析,但如果访问他们的话就可以从本地环境得到更多的信息;(2)在避免重复中间的标记,因为这是已知的解析器;(3)依赖边缘通常是本地的,也就是说,单词和近邻单词之间存在依赖关系。因此,嵌入的培训是用来预测相邻标记的,并且可能会随着学习类似的信息来培训依赖边缘。结果发现,较大的窗口会捕获更多的语义信息,而较小的窗口则更好地反映语法。我们会选择的各自窗口。后来考虑多种来源的语言,但现在假设一个单一的源语言。在 16 种语言数据集中,观察到依赖厘清的 50% 系统会跨越一个词的距离,但 20% 系统会跨越两个词的距离。因此,在一个±1 个字的窗口的 POS 背景下抓住了大部分的依赖关系。算法 1 的步骤 4 查找字嵌入会作为训练神经语言的副作用模型。使用 skip-gram 模型来训练每个上下文的标签字。

2.3 分析算法

在本节中,将展示如何把语法字嵌入到一个分析模型。分析模型是建立在基于相关技术的研究上。他们建立了一个使用基于过度依赖分析器的网络。神经网络分类器将决定该过渡所加载的每个配置。也就是说每个配置所选的列表包括 POS 标签和标签栈,其中队列被提取。每个字 POS 或标签是通过映射层映射到低维向量的。这个层简单地连接了嵌入,然后送入两层神经网络来预测下一个语法分析的动作。对于神经网络的参数集合分类器是字标签的映射层,而对于隐含层是最大输出层。我们通过设置字到句法字嵌入把语法词嵌入到神经网络模型,这在训练期间保持固定,

以便保持跨语言映射。

2.4　模型摘要

解析器适用于资源贫乏的目标语言,开始通过建立源语言和目标语言之间的语法词,具体如算法 1 所示。然后同步使用算法和字嵌入。最后,分析目标语言所使用的这种替代模式。以这种方式,该模型将认识的词项作为目标语言。

3　实验

测试将字嵌入到神经网络上的分析器,这种分析器可以作用于现有的数据集和最新的发布通用依赖树库。之前的工作是使用无标签的不加标点的数据。如果可能的话,还报告标记不加标点。用英语作为这个实验的源语言。这是唯一的训练解析器的结果源语言。如果在更新嵌入解析器培训,这将意味着他们不再嵌入目标语言。

3.1　实验数据

本节将展示所需要的实验数据集。对于语言而言,我们只使用较新的。最重要的是,对于虚化的解析器,通用标记集来映射特定语言标签。句法字的嵌入是使用经过培训的 POS 数据信息。实验有两个基线:第一种个无监督依赖分析器,第二个是虚化分析器。还比较了我们的句法字嵌入与跨语言文字嵌入。交叉语言文字嵌入到解析模型是以同样的方式作为句法词的。这是与很多以前的研究相一致。我们使用直接传输模型。

实验还显示,使用嵌入直接传输的性能模型。我们的模型使用语法词汇嵌入模型。平均来说,最好是1.5% 和1.3% 。对于改善跨语言变化的 COM 与 HB 的嵌入范围落在了0.3% ~2.6% 。这证实了最初的假说,我们需要嵌入字捕获语法而不是语义信息。我们的无监督的依赖解析方法与以前的方法具有不同的资源,即并行数据或类型学的资源。与直接传输的基线模型相比,平均增益 1.5% ,报告的大约 6% 的涨幅。如上所述,对于在这些其他系统中使用的方法,我们的做法是相辅相成的。例如,可以将跨语种词聚类特征并入到我们的模型,或者使用我们的虚化解析器,预计将会导致更好的结果。

3.2　实验与依赖原则

我们也尝试了使用相依树库,在系统中有许多可取的,如依赖类型的属性是相同的跨语言。这消除了目标语言到公共标记集所需要的映射源。其次,不可能在数据所在标签语言之间产生差异。表示在数千个树库语言。第一观察到的是,有些丰富的语言,如捷克语,法语和西班牙语。我们运行了无语法字嵌入与英语所有语言作为源语言的模型,第一个发现是,对于所有的语言,我们的模型使用句法字嵌入外执行是在两个无人机系统之间直接传送的。我们观察到平均提高3.6% (UAS) 和3.1% (LAS)。这种合并的句法字嵌入到模型的一致的改进显示了我们方法的稳健性。这反映了增加标记边缘与未标记的边缘预测只涉及 3 路分类,而标记边缘预测涉及 81 路分类,缩小 UAS 和 LAS 对资源贫乏的语言之间的间隙是基本功夫。在今后的研究领域工作中,在前面的章节所提到的 5 种不同的语言源代码中,我们使用英语作为源语言。然而,英语可能不是最佳选择。对于虚化解析器而言,源语言和目标语言具有类似的句法结构是至关重要的。因此,另一个不同源语言的选择可能会改变性能,如在现有 stud-观察 IES。由于只有 3 转换:SHIFT, LEFT-ARC, RIGHT-ARC。由于依存树库中的 40 个单向通用开关系,每个关系连接到 LEFT-ARC 或 RIGHT-ARC。数 81 来自 1(SHIFT) +40(LEFT-ARC) + 40(RIGHT-ARC)。在通用依存树库 UAS 的评估中。

在本节中,假设我们有多重 PLE 源语言。要了解如何使用不同的源语言的变化时,我们运行最好的模型,在每个语言对依赖树库研究显示,所有的平均源语言的语言有所提升。也考虑了 LAS,但观察到了类似的趋势,因此只报告了 LAS 平均每个源语言。注意,英语是最好的源语言;虽然捷克在 UAS 方面很优秀,但它 LAS 方面效果相对较差。人们可能期望使用不同的源语言所受到影响是由源语料的大小来决定的。通过限制源测试,这就造成了轻微的重在分数,但总体来说使用全尺寸的源语料库还是最好的,以及平均排名由 UAS 和 LAS 保持不变。10 个语言视为属于 5 家庭:(法语,西班牙语,意大利语) 日耳曼(德语,英语,瑞典语),斯拉夫(捷克),乌拉尔(匈牙利,芬兰),以及凯尔特人(爱尔兰)。乍一看,似乎语言在同一个家庭往往表现良好。例如,对于这两个最好源语言法语和意大利语,最好西班牙源语言是法语。然而,这并不是所有目标语言都是如此。对于这两个最好源语言芬兰语和德语。对于适当的源语言的最好选择是没有语言的家庭信息。因此,我们提出了两种方法来预测对于给定的目标语言的最佳源语。在制定这些方法,假设因为做了一个给定的目标语言,不能访问任何分析的源语言数据,这种模型预测则是不成功的。第一种方法是根据詹森-香农分歧词性正贡献克($1<N<6$) 的一

对语言。第二种方法将每个语言转换成二进制的特征向量。

4 结论

在以前的大部分工作中,跨语种转移依赖解析是一直依靠大的平行语料库。然而,并行数据是稀缺资源贫乏语言。在文中,我们的第一部分是建设一个依赖解析器,该解析器的数据并不需要并行数据。改进了使用神经网络的句法词嵌入的虚化解析器。发现,语法词是善于捕捉语法的最好信息,特别适用于依赖解析。与此相反,国家最先进的跨语言依存分析是非监督的,我们方法不依赖于并行数据。虽然国家最先进的方法比我们的方法取得了更大的收益基线,但由于其较低的资源需求,我们的方法可能会更广泛地应用到资源贫乏的语言。

第二部分是研究提高性能的方法是可用性。提出了两种方法用来选择单一源语言,总是选择英语作为改进源语言。然后,我们发现,可以结合所有的源语言信息的来进一步提高性能。综上所述,没有任何并行数据,在普遍依赖树库中,我们设法改善了解析器虚化。在今后的工作中,还可以建立 POS 嵌入和弧标签的使用。这可以帮助系统的跨语言能够更自由地移动,大大提高跨语言方法的实用性。

参考文献:

[1] Wenbin Jiang, Qun Liu. Dependency Parsing and Pro-jection Based on Word-Pair Classification [C]. Pro-ceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL). Uppsala, Swe-den 2010:12-20.

[2] 宣云干. 基于潜在语义分析的社会化标注系统标签语义检索研究[D]. 南京: 南京大学,2011.

[3] 宋晓雷,王素格,李红霞,等. 基于概率潜在语义分析的词汇情感倾向判别[J]. 中文信息学报,2011,25(2): 89-94.

[4] 单斌,李芳. 基于 ELW 话题演化研究方法综述[J]. 中文信息学报,2010,24(6):43-49.

[5] Liu Dong C, Jorge Nocedal. On the limited memory BFGS method for large scale optimization[J]. Mathematical Programming,1989,45(3):503-528.

[6] Liu Lemao, Hailong Cao, Taro Watanabe, et al. Locally training the log-linear model for SMT[J]. In Proceedings of EMNLP/CoNLL,2012,402-411.

[7] Klein D, Manning C D. Corpus-based Induction of Syntactic Structure: Models of Dependency and Constituency[C]. Barcelona, Spain,2004:478-485.

[8] Gulsen Eryigit, Kemal Oflazer. Statistical Dependency Parsing of Turkish[C]. 11th Conference of the Euro-pean Chapter of the Association for Computational Lin-guistics (EACL). Trento, Italy, 2006:89-96.

[9] E. Charniak. A Maximum-entropy-inspired Parser [C]. Proc. NAACL. Seattle, Washington, USA, 2000:1396-1400.

[10] A. Ratnaparkhi. Learning to Parse Natural Language with Maximum Entropy Models[J]. Machine Learn-ing,2002,34:151-175.

[11] Liu Lemao, Liang Huang. Search-aware tuning for machine translation[J]. In Proceedings of EMNLP,2014:1942-1952.

[12] Liu, Lemao, Taro Watanabe, Eiichiro Sumita, et al. Additive neural networks for statistical machine translation[J]. In Proceedings of ACL, 2013:791-801.

[13] Liu, Lemao, Tiejun Zhao, Taro Watanabe, et al. Expected error minimization with ultraconservative update for SMT[J]. In Proceedings of COLING: Posters,2012:723-732.

[14] Liu, Lemao, Tiejun Zhao, Taro Watanabe, et al. Tuning SMT with a large number of features via online feature grouping[J]. In Proceedings of IJCNLP,2013:279-285.

[15] Lo, Chi-kiu, Karteek Addanki, Markus Saers, et al. Improving machine translation by training against an automatic semantic frame based evaluation metric[J]. In Proceedings of ACL: Short Papers,2013:375-381.