数据标注方法比较研究：以依存句法树标注为例

周明月，龚晨，李正华*，张民
苏州大学计算机科学与技术学院，苏州，中国
{20194227026，cgong}@stu.suda.edu.cn，{zhli13，minzhang}@suda.edu.cn

摘要

数据标注最重要的考虑因素是数据的质量和标注代价。我们调研发现自然语言处理领域的数据标注工作通常采用机标人校的标注方法以降低代价；同时，很少有工作严格对比不同标注方法，以探讨标注方法对标注质量和代价的影响。该文借助一个成熟的标注团队，以依存句法数据标注为案例，实验对比了机标人校、双人独立标注、及本文通过融合前两种方法所新提出的人机独立标注方法，得到了一些初步的结论。

关键词： 数据标注方法；人机独立标注；机标人校；多人独立标注

Comparison Study on Data Annotation Approaches: Dependency Tree Annotation as Case Study

Mingyue Zhou, Chen Gong, Zhenghua Li, Min Zhang
School of Computer Science and Technology, Soochow University, Suzhou, China
{20194227026，cgong}@stu.suda.edu.cn，{zhli13，minzhang}@suda.edu.cn

Abstract

The important considerations for data annotation is the quality and the cost of the data. According to the investigation, we find that the data annotation in the field of natural language processing usually adopts the annotation approach of first model annotation and then human correction to reduce the cost. At the same time, there are few efforts to strictly compare different annotation approaches to explore the effects of annotation approaches on annotation quality and cost. With the help of a mature annotation team, this paper takes dependency tree annotation as case study, and experimentally compares three data annotation approaches of first model annotation and then human correction, double-blind annotation , and human-model double-blind annotation proposed by this paper through the fusion of the first two methods, and initially finds some very interesting results.

Keywords: data annotation method , human-model double-blind annotation , first model annotation and then human correction , double-blind annotation

1 引言

*通讯作者

近年来，随着深度学习技术的发展，自然语言处理研究取得了巨大突破。与传统机器学习技术相比，深度学习更加依赖大规模、高质量的有答案数据。一方面，有监督学习方法需要输入数据和对应的答案，以学习模型参数。另一方面，评价阶段也需要已知答案的数据，从而比较和分析不同模型的优劣。数据标注是一个具有很强应用价值的科学研究问题。很多研究者都致力于数据标注工作，通过LDC[0]、ELRA[1]等平台发布数据。然而，目前大多数研究者都仅仅将数据标注的实施过程看成一个繁琐的工程性工作，鲜有学者从方法论的层面出发，通过严格实验对比，分析不同标注方法的优劣。

本文从方法论的层面对不同标注方法进行实验对比，研究如何更有效地开展人工数据标注。根据答案是否唯一，可以将数据标注划分为两种类型：确定性答案的数据标注和开放性答案的数据标注。大部分自然语言处理任务都采用确定性答案的数据标注，如分词、词性标注、命名实体识别、句法分析等。确定性答案的含义是指研究者通过制定标注指南，保证每一个数据只有一个唯一的答案。然而，对于一些问题，如机器翻译和句子复述，由于答案表达的多样化，很难通过标注指南指定唯一答案，因此必须采用开放性答案的数据标注。本文以依存句法数据标注为案例，主要关注确定性答案的数据标注。然而，值得说明的是，本文研究取得的发现和结论对于开放性答案的数据标注也是有价值的。

数据标注主要考虑两方面因素：数据质量和标注代价。因此，不同标注方法的对比，也应该从这两方面入手。具体而言，所谓标注数据质量高，是指标注结果能够准确的区分输入数据，尤其模棱两可的情况，不同标注者也会产生一致的标注结果。我们总结发现标注错误主要有三种来源：1）无规律的明显错误。这种错误通常是由于误操作、注意力不集中等偶然因素，也可能是因为态度不认真或对标注指南不熟悉；2）同一标注人员对标注指南理解发生变化而导致的有规律的不一致错误；3）不同标注人员对标注指南理解不同而产生的不一致错误。后两种错误通常来自于一些模棱两可的困难样本，但也有少数情况下是对于一些比较简单的样本。可以看出，无论是哪种错误，都会给模型训练和评价带来很大的干扰。

除标注质量外，人工数据标注的另一项重要考虑因素是标注代价，具体包括人力、金钱和时间代价。质量和代价这两个因素通常是矛盾的。多个人独立标注同一个任务，进而由能力更强的标注人员对不一致的结果进行审核，是最直接、有效的提高质量的做法，但是也会显著增大标注代价。当然，研究者发现利用主动学习和局部标注技术，可以在不影响质量的前提下，尽可能降低标注代价 (Settles, 2009; Olsson, 2009; Li et al., 2016b)。受篇幅所限，本文对此不作深入探讨。

常用的数据标注方法包括：人标人校、机标人校和多人独立标注。其中，机标人校是一种最典型的人机协作方式，可以有效降低标注代价。为此，大多数数据标注工作均采用机标人校进行大规模标注，如宾州英文树库 (Marcus et al., 1993)、宾州中文树库 (Xue et al., 2005)、SINICA语料库 (Chen et al., 1996)、北大多视图汉语树库 (邱立坤et al., 2015)、北大人民日报词法数据 (俞士汶et al., 2002)、清华汉语树库 (周强et al., 2002)。人标人校和多人独立标注方法通常在测试数据标注或项目初始阶段时采用 (Xia et al., 2000; McDonald et al., 2013)，后者也常用于计算一致性 (Xue et al., 2005; Kessler et al., 2010)。本文在第2节详细讨论相关工作。进而，本文还提出一种新的人机协同标注方法，可以有效融合机标人校和多人独立标注方法的优势。我们将这种标注方法命名为人机独立标注方法。

我们的调研还发现目前很少有工作对不同标注方法从数据质量和标注代价两个角度进行严格实验对比。宾州树库 (Marcus et al., 1993) 构建过程中，研究者通过很小规模的标注实验，对机标人校的效率进行了初步的探索。基于过去几年来一直从事依存句法树标注的研究基础，我们借助成熟的标注团队，以依存句法树标注为案例，尝试通过严格实验，对比不同标注方法对质量和代价的影响。

依存句法分析的目标是给定输入句子，构建一棵依存句法树，捕捉句子内部词语之间的修饰或搭配关系，从而刻画句子的句法和语义结构 (Kübler et al., 2009)。图1为一棵依存句法树的示例。其中，$是一个伪词，指向句子根节点。作为依存树的最基本单元，一条依存弧包含三要素：核心词（父亲）、修饰词（儿子）和依存关系标签。例如，“奶奶$\xleftarrow{\text{subj}}$打转”这条依存弧表示“打转”为核心词，“奶奶”为修饰词，依存关系标签为subj（主语）。

具体而言，本文的主要贡献可以总结如下：

[0] https://www.ldc.upenn.edu/
[1] http://www.elra.info/

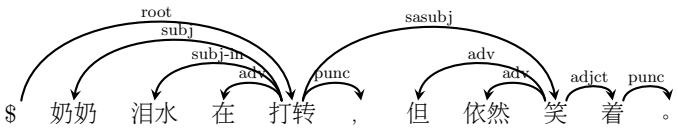


Figure 1: 依存句法树示例

- 我们对目前已有的数据标注工作进行了深入调研，总结出几种常用的数据标注方法。
- 我们提出一种新的人机协作标注方法，并命名为人机独立标注，以结合不同标注方法的优势。
- 我们通过严格实验，比较了机标人校、双人独立标注、人机独立标注三种典型的标注方法。对于质量和成本的影响，取得了一些初步的结论，希望可以帮助从事数据标注工作的研究者。

2 相关工作调研和分析

本节调研并总结了数据标注方法方面的相关工作。由于数据标注工作的范围非常广，我们的调研主要关注词法句法语义分析、信息抽取等确定性答案的数据标注工作，而不涉及机器翻译、对话等开放性答案的数据标注工作。

自从20世纪90年代，自然语言处理研究从规则系统时代进入到统计模型时代，标注语料库资源建设就变成很多研究工作必不可少的基础。深度学习时代，神经网络模型可以灵活设置参数规模，因此对于有标注数据的需求更大。除了学术界，工业界常常遇到一些具体的问题，需要从无到有标注数据以训练模型。因此，对数据标注方法进行系统研究，是一件兼具学术和实践价值的工作。通过调研，我们发现人标人校、机标人校、多人独立标注是三种最常用的数据标注方法。其中，多人独立标注方法获得的数据质量最高，但是标注代价也最大。机标人校的代价最小，但是质量通常也最低。

人标人校的方法. ACE（The Automatic Content Extraction）数据集 (Strassel and Mitchell, 2003)主要目的是支持多语言信息抽取研究，涵盖了实体识别、事件抽取等任务，其主要采用了人标人校的标注方法。在自然语言处理领域的标注项目中，采用人标人校方法的案例较少。我们猜测主要原因是人标人校方法存在两个问题。第一是标注者的精力浪费问题。对于密集操作型的标注任务，即标注一个数据需要很多操作，标注者的大部分精力都花在鼠标点击或键盘输入等操作上，而实际分析和理解数据所占的精力较少。例如，分词任务要求标注者在输入句子中加入分隔符，从而产生词序列。类似的，词性标注任务、句法标注任务、语义标注任务等均属于密集操作型任务。

人标人校方法的第二个问题是校对者的认同倾向。由于人的思维惰性，校对者往往会倾向于认同标注者的结果，只会修改明显的错误，因此校对阶段的错误纠正率，即纠正的错误数相比于真实的错误数，通常会比较低。对于密集操作型的任务，如分词、词性标注、依存句法树标注等，由于一个数据包含很多标注信息，校对者更容易遗漏标注错误。当然，理想情况下，我们希望校对者比标注者的标注能力更强，从而提高错误纠正率。但是需要考虑的是，很多情况下，能力明显更胜一筹的标注者的比例不会很高，可能无法应付校对工作量。

机标人校方法. 机标人校方法是一种常见的人机协作标注方法。相比人标人校方法，机标人校方法可以解决标注者的精力浪费问题。首先，机器在数据上产生自动结果，然后由校对者定位并修改结果中的错误。对于密集操作型的标注任务，机器的大部分结果很可能都是对的，因此只需要校对者进行少量的操作，就可以完成标注。并且，随着标注数据的增多，可以重新训练模型，不断提高机器的性能。

大多数标注数据集采用机标人校方法进行大规模标注。宾州树库（Penn Treebank, PTB） (Marcus et al., 1993) 对超过450万单词的英文语料进行了词性和短语结构句法树标注。项目初期，他们采用了第三方自动标注工具，产生词性和句法结果；然后对不符合PTB标注指南的结果，基于规则匹配并自动转化；最终由语言学家进行校对，确定答案。在获得一定量的标注数据之后，他们将其作为训练数据训练了统计模型，继续以机标人校的方式扩大语料

库规模。类似的，Sinica中文句法树库 (Chen et al., 1996)、北大词法数据 (俞士汶et al., 2002)、清华汉语树库 (周强et al., 2002)、宾州中文树库（Penn Chinese Treebank, CTB） (Xue et al., 2005)、北大多视图句法树库 (邱立坤et al., 2015)均采用机标人校的标注方法。

机标人校方法可能存在冷启动问题，即项目初期没有任何标注数据以训练模型。这种情况下，就需要用一些简单的规则来产生结果。PTB项目中的做法值得参考，即利用相关资源得到初步结果，然后根据标注指南进行规则转换。

机标人校的方法同样存在校对者的认同倾向问题，即校对者倾向于认为机器的结果是正确的，从而导致纠错率较低。由于标注得到的语料会继续用来训练模型，因此修正的错误会越来越少。这样就会导致一个更严重的面向模型收敛的问题，即整个数据标注项目的目标会转变为快速提高模型准确率，而非标注出高质量的数据。一个好的标注项目应该是面向问题收敛，即通过标注指南精确刻画问题，严格按照标注指南区分不同的数据。

多人独立标注. 多个标注者对同一个数据进行独立标注，标注过程中不会看到别人的结果，因此可以从根源上解决认同倾向问题。如果多个标注者结果相同，那么就作为最终答案；否则，则通过某种方式对多个结果进行对比，并确定最终答案。多人独立标注在数据标注实践中也被广泛使用，例如捷克语依存树库 (Hajic et al., 2001)、CAMR (Li et al., 2016a)、阿姆斯特丹隐喻语料库 (Ide and Pustejovsky, 2017)、苏州大学汉语开放依存树库 (郭丽娟et al., 2019)。多人独立标注的明显优势是可以发掘对问题的理解差异，促进标注指南的完善、标注者水平的提高等。但是，很显然，多人独立标注的标注代价非常大。因此，研究者通常在三种场景下采用多人独立标注方法：1）项目初期 (Ševčíková et al., 2007; McDonald et al., 2013)；2）标注测试集数据时 (Xia et al., 2000)；3）计算标注一致性时 (Xue et al., 2005; Kessler et al., 2010)。

捷克语依存树库（Prague Dependency Treebank, PDT））项目 (Hajic et al., 2001)对180万单词的捷克语料标注了形态、词性和句法结构。在形态学层，每个样本由两个标注人员独立标注，两个标注不一致的样本由第三位标注人员进行处理。而句法层的标注，由于句法结构非常复杂，同时作者的经验有限，因此他们迭代地进行制定、修改标注指南和标注数据。具体采用什么标注方法进行句法标注，文章没有给出明确说明。另外值得一提的是，PDT标注项目构造并大量使用工具，用来自动检测和纠正错误。

苏州大学发布的汉语开放依存树库（Chinese Open Dependency Treebank, CODT） (郭丽娟et al., 2019)主要采用双人独立标注方法。同一个数据由两位标注人员分别进行独立标注。对于不一致的结果，由第三位标注水平更好的标注人员（审核专家），进行对比并确定最终答案。为了帮助标注者提高水平，他们还提出一种学习机制，即标注系统会将错误的标注反馈给标注人员，并强制要求标注人员根据正确答案亲自纠正错误。进一步，为了发现审核专家错误，标注人员学习过程中，可以对审核专家的错误进行投诉。投诉问题会推送给标注指南制定者（权威专家）解决。

多人独立标注和机标人校的结合. 对于密集操作型任务，多人独立标注存在非常严重的精力浪费问题。一个很自然的解决思路是和机标人校方法相结合。德语TIGER树库包含3.5万句德语新闻句法标注语料 (Brants et al., 2002)。标注过程中，他们首先用自动分析器生成结果。进而，每个句子由两位标注人员独立标注，即纠正机器的错误。对于不一致的结果，两个标注者会进行讨论，形成唯一答案。

这种融合方法仍然存在认同倾向问题，并导致面向模型收敛的问题。也许多个不同的机器学习模型，将不同模型的结果给不同的标注人员，形成差异化，可以从一定程度上缓解这一问题。

与TIGER项目不同，本文提出一种新的融合机标人校和多人独立的标注方法，即人机独立标注方法，可以从根本上解决认同倾向问题。

不同标注方法的对比研究. 在数据标注研究工作中，极少看到学者对不同标注方法进行严格对比，以探讨标注方法对标注质量和代价的影响。很显然，客观、公平比较多个标注方法，并取得具有较普遍适用性的结论，是一项代价很高的工作。

宾州树库在开展词性标注任务的早期，为了平衡标注速度、标注人员一致性和准确率，实验对比了人工标注和机标人校 (Marcus et al., 1993)。其实验结果表明人工标注比机标人校标注时间长大约两倍，标注人员不一致率高两倍，错误率约高50%。宾州中文树库也通过类似实验，说明在句法标注任务上，机标人校的标注速度显著高于完全的人工标注 (Xue et al.,

2005)。但是，这两个工作是对人机协同标注方法的早期探索，对实验的细节介绍都很简略，并且规模非常小。

由于缺乏方法层面的比较研究，数据标注项目在选择标注方法时缺少可靠的参考信息。本文借助一个成熟的标注团队，以依存句法数据标注为案例，对机标人校和多人独立标注这两种常用的标注方法，以及我们提出的人机独立标注，从质量和代价两方面，进行了深入对比，并取得了一些初步结论。

3 标注方法

本文重点比较和分析了三种数据标注方法，即机标人校、多人独立标注、以及本文提出的人机独立标注。其中，机标人校方法被很多数据标注项目采用，是一种主流的标注方法，而多人独立标注则在计算标注一致性和产生高质量的评价数据时被广泛采用。本文提出的人机独立标注方法是对机标人校和多人独立标注方法的一种有效融合。本节详细介绍了三种标注方法的具体实施。

3.1 机标人校

机标人校是一种常见的人机协作标注方法，是指由机器扮演标注人员的角色在数据上产生自动标注结果，然后人工校对机器自动生成的标注结果。所谓机器，是指在已有标注数据上训练的机器学习模型。机标人校的优势在于，对密集操作型的标注任务，机器自动标注的结果可能大部分是正确的，因此校对者只需要进行少量的操作就可以完成标注，相较于完全的人工标注，节约了大量的人力和时间。并且，随着标注数据的增加，可以重新训练模型，不断提高机器的性能，反过来进一步提高标注工作的效率。

本文实验中采用的机标人校如图2a所示，对一个待标注样本，首先由模型产生自动标注结果，然后交给标注人员进行校对，标注人员如果发现错误就进行相应修改。因为人机独立标注方法中为了确定唯一答案，含有审核过程，为了提高二者的可比性，在机标人校方法中增加二次校对。二次校对由更有经验、准确率更高的标注人员进行。

校对者的认同倾向是机标人校的固有问题，即校对者倾向于认为机器的结果是正确的，从而导致纠错率较低。标注得到的语料会继续用来训练模型，因此修正的错误会越来越少，这将导致标注数据面向模型收敛，而不是面向标注任务收敛。多次校对可以弥补部分因认同倾向而导致的疏忽，但也增加了时间和金钱成本，同时也可能另外引入标注结果的不一致性。

3.2 双人独立标注

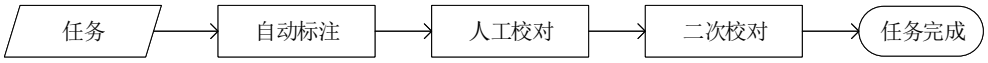
本文采用的双人独立标注是多人独立标注的一种常见形式。多人独立标注是指同一个任务由多个标注人员独立标注。独立标注是指标注人员之间不进行讨论，从而最大限度地挖掘不一致。如果多人标注的结果一致，我们通常认为标注答案正确，任务完成，如果结果不一致，则需要通过某种方式确定最终答案，常采用审核或投票的方式。多人独立标注可以从根源上解决标注者的认同倾向问题，发掘对标注任务的理解差异，促进标注指南的完善、标注者水平的提高等。但多人独立标注的标注代价较大，尤其是金钱成本，随着标注人数的增加而成比上涨。

本文实验中的双人独立标注如图2b所示，每个待标注任务由两个标注人员独立标注，如果答案一致，则任务完成，如果答案不一致，由经验更丰富的标注人员进行审核，确定最终答案。

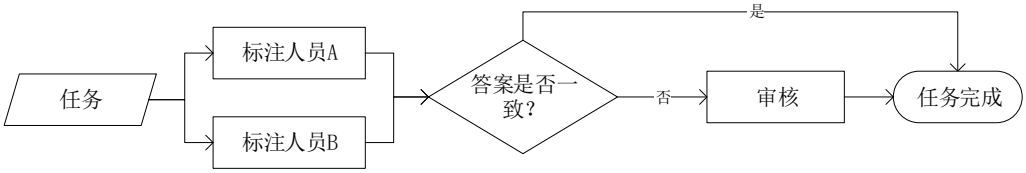
3.3 人机独立标注

基于对机标人校和多人独立标注的分析，如何能结合两者优势、降低成本的同时解决认同倾向，是我们自然而然想到的问题。本文提出一种新的人机协同的标注方法，即人机独立标注，用模型自动标注的答案作为待标注任务的一个标注答案，再由标注人员独立给出另一个标注答案，独立标注说明标注人员无从得知模型标注的答案。如图2c所示，如果标注人员和模型给出的标注不一致，则交由审核专家审核确定，如果标注一致，则任务结束。

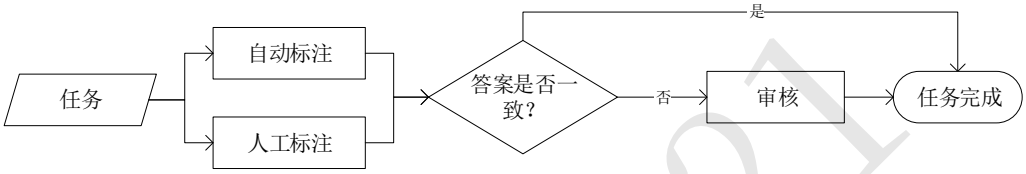
人机独立标注方法类似双人独立标注方法，因此同样也能从根源上解决标注者的认同倾向问题。并且，将其中近一半的人力替换成机器，自然属于人机协同的标注方法，具有借助机器减少标注成本的特点。这样一来，直觉上，人机独立标注兼具机标人校和双人独立标注的优势。



(a) 机标人校



(b) 双人独立



(c) 人机独立

Figure 2: 本文重点关注的三种标注方法

此外，人机独立标注不仅限于本文实验所采用的形式。它不仅可以是一人一机，也可以是一人多机、多人一机或多人多机。使用多个机器自动标注答案，可以有效利用不同机器学习模型各自的优势，挖掘不一致性，为数据分析提供多样化的信息。对质量要求相对不高的标注项目，在后续的审核过程中也可以用多数投票确定最终答案，以此进一步节省成本。在开放性答案的标注任务或众包任务中，也具备应用前景。

4 数据

4.1 数据来源

数据标注实验涉及人工劳动，因此代价比较大。同时，我们希望数据标注实验的结果能够符合实际的项目需求，从而具有更大的价值。最终，我们选择了两个来源的数据。第一个来源为北语句法结构树库1.0 (卢露et al., 2020)，包含了百度百科、新浪和新华社新闻、国家专利等文本。第二个来源为机器翻译中英对齐文本NIST（MT02-MT06）(Zhang et al., 2019)，文本主要来自法新社和新华社新闻。

本文实验重点比较机标人校、双人独立、人机独立三种标注方法。但是，为了控制实验的复杂度，我们以人机独立为枢轴，首先在第一个来源的数据上对比机标人校和人机独立方法，然后在第二个来源的数据上对比双人独立和人机独立方法。

4.2 数据处理

在正式标注开始之前，我们对数据进行一些预处理、选取和划分。我们首先对数据进行了去重、过滤特殊符号、限制句子长度、全角字符转成半角字符的预处理操作。然后我们使用自动分词工具将数据分词，输入训练好的句法模型，得到自动标注答案。

为了探索在随机数据和困难数据上标注方法效果的差异，我们随机选取部分待标注数据，并通过计算待标注词难度来选择困难待标注数据。Li et al. (2016b)研究了句法分析任务中，基于局部标注数据的主动学习方法，取得了令人满意的结果。主动学习是指让模型主动选择数

数据批次	来源	句子长度（词数）	局部标注比例	难度分布		
				难度	句子数	标注词数/总词数
BCC	北语句法结构树库	[5,40]	20%	随机	2125	8454/35658
				困难	1901	6314/32006
MT	NIST(MT02-MT06)	[5,40]	20%	随机	2413	10280/44189
				困难	1914	8282/35908

Table 1: 数据分布

据，以最大程度减少标注代价。借鉴Li et al. (2016b)的工作，我们在句法模型输出的自动标注结果中，保留了每条弧的边缘概率作为置信度来评价待标注词的标注难度，即认为置信度越低的待标注词的标注难度越大。同样地，我们也使用局部标注方法，意味着标注者只需要标注句子中部分词语。假设局部标注比例为20%，那么图1中句子一共8个词，只选择其中2个词作为待标注词。基于置信度的计算，我们可以对数据进行随机采样和困难采样。随机采样是指从预处理的全部未标注数据中抽取一定数量的句子，这些句子的待标注词是按局部标注比例随机选择的。在随机采样之后，从剩余的未标注词中进行困难采样，按局部标注比例选择句子中置信度最低的词作为待标注词。

最后，为了避免重复劳动、减少待标注数据与句法模型训练数据的相似性，我们对随机采样数据和困难采样数据都进行了相似度过滤，按设定阈值去除自相似度过高或与训练数据相似度过高的待标注数据。

4.3 数据分布

经过数据处理，我们获得了两批待标注数据，数据分布如表1所示。两批数据的句子长度都限制在5-40个词，局部标注比例为20%，即标注词数占总词数的20%。其中BCC批次来源于北语句法结构树库[2]，随机采样数据为2125句，困难采样数据为1901句。MT批次数据来源于机器翻译中英对齐文本NIST（MT02-MT06），分为随机采样数据2413句和困难采样数据1914句。

5 实验设置

依经验推测，机标人校标注速度快、金钱成本低，但标注质量相对较低，且有标注者的认同倾向问题。而多人独立标注虽然从根本上解决了认同倾向问题，能够大大提高标注质量，但其标注成本也远远大于机标人校。直觉上，融合了前述两种方法的人机独立标注应该兼具机标人校和多人独立标注的优势。为了验证这点，我们以人机独立标注为枢轴设计了两组实验。

第一组实验在BCC数据上对比机标人校和人机独立标注。BCC数据包含随机采样数据和困难采样数据，将随机采样数据分为平均的两部分，一部分使用机标人校标注方法进行标注，另一部分使用人机独立标注。困难采样数据也采取同样的策略。

第二组实验在MT数据上对比人机独立标注和双人独立标注。与第一组实验一样，MT数据的随机采样数据和困难采样数据也分别平均划分为两部分，分别使用人机独立标注方法和双人独立标注方法完成标注实践。

5.1 标注工具和团队

为了支持这两组实验，我们使用苏州大学NLP标注系统[3]作为线上即时标注工具，将每一个依存句法标注任务都标注成如图1所示的句法树。苏州大学NLP标注系统可以对自然语言处理领域的多项标注任务进行标注，例如依存句法、语义角色标注、命名实体识别等。它支持机标人校、双人独立标注等标注方法，并且标注方法对标注人员不透明，标注人员只需要关注如何标注当前任务，而不知道当前任务处于哪种标注方法。该标注工具通过简单地设置数据输入格式，就可以随机交替使用不同标注方法进行数据标注。借助这个标注工具，我们可以方便、快速地完成标注实践。

除了合适的标注工具之外，我们还借助一个成熟的依存句法标注团队开展标注工作。该标注团队中包括10-15位具有长期标注依存句法经验的标注人员和5位平均准确率在90%以上的审

[2]http://bcc.blcu.edu.cn/

[3]http://139.224.234.18/anno-sys/index.php

核专家。标注人员和审核专家都是经过标注培训的本科学生或研究生，具有计算机科学和中文母语的知识背景。

5.2 评估方法

- **质量评估.** 为了评估通过不同标注方法得到的数据的标注准确率，我们采用抽样重标的方法。具体而言，对已经使用机标人校、双人独立标注或人机独立标注方法标注过的数据，随机抽取其中20%的句子，使用双盲标注方法重新标注。需要注意的是，使用机标人校、双人独立标注或人机独立标注得到的首次标注答案将作为重标时双盲标注中的一个标注答案与重标时的人工标注答案进行不一致审核。重新标注后获得的结果为最终的正确结果，据此计算第一次标注结果的准确率。除了准确率以外，我们还计算了标注人员的一致性，分为弧一致性和句子一致性。三个评价指标的公式分别是：

$$\text{准确率} = \frac{\text{第一次标注正确的词数}}{\text{总标注词数}} \tag{1}$$

$$\text{弧一致性} = \frac{\text{两个独立标注答案一致的弧数}}{\text{总弧数}} \tag{2}$$

$$\text{句子一致性} = \frac{\text{两个独立标注答案一致的句子数}}{\text{总句子数}} \tag{3}$$

其中，标注正确的词是指弧和依存关系标签都标注正确，标注答案一致的弧是指弧和依存关系标签都一致。

- **时间成本.** 标注工具记录了标注人员每个句子的标注时间，我们将所有标注人员的标注时间总和除以标注的句子数，分别计算了平均标注时间、平均审核时间和平均总时间。其中机标人校的审核时间即为二次校对的时间成本。为了减少标注人员因故挂机的影响，标注工具限制在一个任务上停留的时间为五分钟以内。在标注人员较多、标注量比较大的情况下，标注人员的主观因素对标注时间的影响会比较小。

- **金钱成本.** 金钱成本分为标注成本和审核成本。标注成本只计算标注人员进行标注时的工资。为了更接近实际标注项目中对标注人员标注水平的考察和奖励机制，将标注人员的准确率视为计算工资的重要系数，因此单个标注人员的工资计算公式为：工资 = 弧单价 × 弧数 × 准确率^2。对所有标注人员的标注工资求和再除以总弧数，得到实际标注弧单价。审核成本只计算审核人员进行审核时的工资。只有两个标注答案不一致的句子需要进行审核，并且只计入审核人员纠正的弧数。审核工作需要更有经验的人员进行，花费的时间也更多，因此审核弧单价是标注弧单价的两倍。单个审核人员的工资计算公式为：工资 = 弧单价 × 弧数。对所有审核人员的审核工资求和之后除以总弧数，得到实际审核单价。实际标注单价和实际审核单价相加即为实际总单价。

6 实验结果

6.1 机标人校与人机独立标注的比较实验

我们在BCC数据上进行了机标人校与人机独立标注的实验，结果如表2所示，我们从以下三个方面进行分析。

准确率：在随机采样数据上，机标人校的准确率达到94.1%，如果进行二次校对，准确率提高到95.5%，而人机独立标注准确率为97%。在困难样本上，机标人校的准确率显著降低，只有83.4%，二次校对提高到89.3%，而人机独立标注则达到了91%。在随机采样数据上，两种标注方法的差距尚且为1.5%-2.9%，在困难采样数据上，差距则增加到了1.7%-7.4%。可见，机标人校准确率显著低于人机独立标注，在困难样本上差距进一步增大。虽然二次校对能大幅提

	困难			随机		
	机标人校		人机独立标注	机标人校		人机独立标注
	无审核	有审核		无审核	有审核	
抽样准确率(%)	83.4	89.3	91.0	94.1	95.5	97
句子数	953		948	1062		1063
弧一致性(%)	71.71		63.18	91.50		87.28
句子一致性(%)	41.76		26.27	75.42		64.44
单位总时间(秒/句)	82		87	55		66
单位标注时间(秒/句)	56		61	38		56
单位审核时间(秒/句)	26		26	17		10
总单价(元/弧)	1.47		1.37	1.13		1.06
标注单价(元/弧)	0.75		0.61	0.93		0.81
审核单价(元/弧)	0.72		0.76	0.20		0.25

Table 2: 机标人校与人机独立标注的比较结果

高准确率，但仍然不及人机独立标注。其原因可能正是机标人校的标注者的认同倾向问题，除了标注人员标注水平而导致的标注错误之外，标注人员可能会受到机标答案的影响或由于任务的枯燥而产生惰性，因而遗漏错误的机标结果。而人机独立标注时，更需要标注人员专注于任务，独立思考给出答案，更易发掘机标的错误、提高标注质量。

一致性：在机标人校和人机独立标注中，一致性计算的是机标答案和人标答案的一致性。结果显示，无论是弧一致性还是句子一致性，机标人校的一致性显著高于人机独立标注，说明机标人校方法中标注者的认同倾向问题确实存在。

标注成本：分为两部分：时间成本和金钱成本。模型自动标注的时间可以忽略不计，因此时间成本只计入人工消耗的。结果显示，机标人校的时间成本显著低于人机独立标注，即便进行二次校对，其花费的时间也低于人机独立标注。产生差距的原因有两个，其一是在标注过程中，人机独立标注方法中标注人员需要更多时间独立思考，而不是对已有答案做快速判断。其二，对于密集操作型的标注任务，在人机独立标注中标注人员的操作行为比机标人校更多。从金钱成本看，一次校对的机标人校的成本明显低于人机独立标注，但加上二次校对之后，其成本就略高于人机独立标注了。其原因可能是二次校对时也存在认同倾向问题，导致标注人员的准确率虚高，拉高了实际标注单价。

6.2 双人独立标注与人机独立标注的比较实验

我们在MT数据上进行了人机独立标注与双人独立标注的比较实验，结果如表3 所示，我们分以下三个方面具体分析。

准确率：在随机采样数据上，人机独立标注的准确率达到93.6%，低于双人独立标注的94.1%，但差距仅为0.5%。而在困难采样数据上，人机独立标注准确率为85%，双人独立标注准确率为89.2%，差距增加到4.2%。双人独立标注的标注质量整体高于人机独立标注，在困难采样数据上差距更明显。其原因可能是：1）人类对自然语言标注任务的理解更为灵活，经验更丰富，人机独立标注相当于把其中一个人工替换成泛化能力低于人类、经验更少的标注者，因此对标注质量造成了负面影响。2）困难数据常常包含非常规的灵活语序或口语化表达，由有限规模的数据训练的自动标注模型能够理解的语言信息范围远小于人类，对困难数据给出的答案其随机性更高。继续扩大训练语料的规模、增加其包含的语言现象，能逐渐缩小自动标注模型与人工标注的差距。

一致性：双人独立标注的一致性总体高于人机独立标注。在随机采样数据上，二者差距很小，说明对于随机数据，自动标注水平与人工标注水平相当。但在困难数据上，差距显著增大，说明自动标注模型对困难样本的理解与人工标注者相差很大，可能是因为自动标注模型对困难数据给出的答案随机性更高。

标注成本：从时间成本看，人机独立标注的标注速度比双人独立标注快了一倍，其中主要原因是自动标注替代了一半的人工标注。从金钱成本看，双人独立标注的标注成本是人机独立标注的两倍左右。审核成本在随机采样数据上相差不大，在困难采样数据上，由于人机独立标注的一致性降低的幅度更大，其审核成本明显高于双人独立标注。但二者相加之后，双人独立标注的总金钱成本仍然高于人机独立标注。

	困难		随机	
	双人独立标注	人机独立标注	双人独立标注	人机独立标注
抽样准确率(%)	89.2	85.0	94.1	93.6
句子数	967	947	1203	1210
弧一致性(%)	60.47	46.62	77.82	78.22
句子一致性(%)	24.30	11.93	44.97	43.80
单位总时间(秒/句)	152	81	119	55
单位标注时间(秒/句)	121	47	98	37
单位审核时间(秒/句)	31	34	21	18
总单价(元/弧)	1.85	1.75	1.88	1.23
标注单价(元/弧)	1.01	0.57	1.41	0.77
审核单价(元/弧)	0.83	1.18	0.47	0.46

Table 3: 双人独立标注与人机独立标注的比较结果

6.3 综合分析

简而言之，一次校对的机标人校的标注质量显著低于人机独立标注，但标注速度快，标注成本减少将近一半。在机标人校的基础上进行二次校对能有效提高质量，但仍然不及人机独立标注的标注质量，同时其标注成本增加，不低于人机独立标注。而双人独立标注的标注质量较人机标注方法有明显优势，但标注速度慢一半、成本更高。

总的来说，人机协同标注方法的主要优势是速度快、成本低，但其标注质量不如双人独立标注，尤其是在困难数据上，质量差距增大。而本文所提出的人机独立标注方法，作为人机协同标注方法的一种新形式，兼具了双人独立标注和机标人校的优势，取得了一个更为折中的方案。它能够避免标注者的认同倾向问题，在相对于机标人校少量增加标注时间成本和金钱成本的情况下，有效提高标注质量，甚至达到双人独立标注的水平。

7 结论与展望

本文以依存句法树标注为例介绍了我们对数据标注方法的比较研究。我们借助成熟的标注规范、团队和工具，对标注实践中最常用的机标人校方法、双人独立标注方法以及本文提出的人机独立标注方法展开实践实验，对比标注质量和标注成本，得到了结论：人机协同数据标注方法虽然大大减少了标注的时间和经济成本，但其标注质量也低于双人独立标注，并且，在困难样本上，人机协同的准确率更低。更进一步地，同样为人机协同的方法，人机独立标注相较于机标人校，在少量增加标注时间成本和薪资成本的情况下，有效提高了标注质量。

通过本文工作，我们在数据标注、语料库构建方面对标注方法的选择给出了建设性的参考。本文研究内容虽然是机标人校、人机独立标注和双人独立标注，但其意义是普遍的。首先，通过在随机样本和困难样本上的分别对比，我们可以推广结论到简单的标注任务和困难标注任务。对于简单的标注任务，如分类标注、意图识别等，相当于本文实验的随机样本，使用人机协同的方式可以在不过分损害标注质量的情况下有效提高标注速度、降低标注成本，并且人机独立标注方法的质量高于机标人校。对于困难标注任务，如句法标注、语义标注等，相当于本文实验的困难样本，人机协同方式的标注质量会显著降低，因此应尽量采用多人独立标注的方式以保证高质量。其次，本文对人机独立标注的实验设置是双盲标注，即一个样本只有两个标注答案，一个是自动标注答案，一个是人工标注答案。对人机独立标注方法进行扩展，可以建立多模型人机独立标注的方式，对一个样本，给出多个不同标注器或标注人员的标注答案，采用多数投票或进一步审核的方式确定最终答案，同时结合了人机协同的标注速度快、成本低和多人独立标注质量高的优点。最后，本文工作使用基本的管理手段——如工作要求、教学与测试、奖惩机制、标注工具的记录等——来减少标注人员的主观因素和工作态度的影响，没有深入的分析，因为在线上标注的情况下，标注人员的主观因素和工作态度很难规范和监督。因此，我们计划未来研究工作包括：各个标注方法在简单标注任务和困难标注任务上的对比实验；多模型人机独立标注方法的研究实验；标注人员主观因素和工作态度的度量。

致谢

本文的工作受到国家自然科学基金(No. 61876116)和江苏高校优势学科建设工程资助项目支持。感谢匿名评审专家对我们工作提出的建设性修改意见。感谢所有标注人员的参与。

参考文献

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The tiger treebank. In *Proceedings of the workshop on treebanks and linguistic theories*, volume 168.

Keh-Jiann Chen, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu. 1996. Sinica corpus: Design methodology for balanced corpora. In *Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation*, pages 167–176.

Jan Hajic, Barbora Vidová-Hladká, and Petr Pajas. 2001. The prague dependency treebank: Annotation structure and support. In *Proceedings of the IRCS Workshop on Linguistic Databases*, pages 105–114.

Nancy Ide and James Pustejovsky. 2017. *Handbook of linguistic annotation*. Springer.

Jason S Kessler, Miriam Eckert, Lyndsie Clark, and Nicolas Nicolov. 2010. The icwsm 2010 jdpa sentiment corpus for the automotive domain. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC)*. Citeseer.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. Dependency parsing. *Synthesis lectures on human language technologies*, 1(1):1–127.

B. Li, Yuan Wen, Weiguang Qu, Lijun Bu, and N. Xue. 2016a. Annotating the little prince with chinese amrs. In *LAW@ACL*.

Zhenghua Li, Min Zhang, Yue Zhang, Zhanyi Liu, Wenliang Chen, Hua Wu, and Haifeng Wang. 2016b. Active learning for dependency parsing with partial annotation. In *Proceedings of ACL*.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Fredrik Olsson. 2009. A literature survey of active machine learning in the context of natural language processing.

Burr Settles. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Magda Ševčíková, Zdeněk Žabokrtskỳ, and Oldřich Krza. 2007. Named entities in czech: annotating data and developing ne tagger. In *International Conference on Text, Speech and Dialogue*, pages 188–195. Springer.

Stephanie Strassel and Alexis Mitchell. 2003. Multilingual resources for entity extraction. In *Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition*, pages 49–56.

Fei Xia, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch, and Mitchell P Marcus. 2000. Developing guidelines and ensuring consistency for chinese text annotation. In *LREC*.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(02):207–238.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. *arXiv preprint arXiv:1906.02448*.

俞士汶, 段慧明, 朱学锋, and 孙斌. 2002. 北京大学现代汉语语料库基本加工规范. 中文信息学报, 16(5):51–66.

卢露, 矫红岩, 李梦, and 荀恩东. 2020. 基于篇章的汉语句法结构树库构建. 自动化学报, 46:1–11.

周强, 任海波, and 孙茂松. 2002. 分阶段构建汉语树库. In *Proc. of The Second China-Japan Natural Language Processing Joint Research Promotion Conference*, pages 189–197. Beijing, China. p189-197.

邱立坤, 金澎, and 王厚峰. 2015. 基于依存语法构建多视图汉语树库. 中文信息学报, 29(3):9–15.

郭丽娟, 彭雪, 李正华, and 张民. 2019. 面向多领域多来源文本的汉语依存句法树库构建. 中文信息学报, 33(2):34–42.