基于网络词典的现代汉语词义消歧数据集构建

严福康、章岳、李正华*
苏州大学计算机科学与技术学院，江苏省，苏州市
20215227039@stu.suda.edu.cn; hillzhang1999@qq.com; zhli13@suda.edu.cn

摘要

词义消歧作为自然语言处理最经典的任务之一，旨在识别多义词在给定上下文中的正确词义。相比英文，中文的一词多义现象更普遍，然而当前公开发布的汉语词义消歧数据集很少。本文爬取并融合了两个公开的网络词典，并从中筛选1,083个词语和相关义项作为待标注对象。进而，从网络数据及专业语料中为抽取相关句子。最后，以多人标注、专家审核的方式进行了人工标注。数据集[1]包含将近2万个句子，即每个词平均对应约20个句子。本文将数据集划分为训练集、验证集和测试集，对多种模型进行实验对比。

关键词： 数据集；词义标注；词义消歧；网络资源

Construction of a Modern Chinese Word Sense Dataset Based on Online Dictionaries

Fukang Yan, Yue Zhang and Zhenghua Li
School of Computer Science and Technology, Soochow University, Suzhou, China
20215227039@stu.suda.edu.cn; hillzhang1999@qq.com; zhli13@suda.edu.cn

Abstract

The task of word sense disambiguation (WSD) is one of the most classic tasks in natural language processing, aiming to identify the accurate sense of polysemous words in a given context. In Chinese, the phenomenon of polysemy is more prevalent compared to English. However, there is a lack of publicly available Chinese word sense dataset. In this paper, we crawled and integrated two publicly accessible online dictionaries, from which we selected 1,083 words and their corresponding senses for annotation. Additionally, we extracted relevant sentences from web data and specialized corpora. Finally, a manual annotation process was conducted through multi-annotator labeling and expert review. The dataset comprises nearly 20,000 sentences, averaging around 20 sentences per word. We divided the dataset into training, validation, and testing sets, and conducted experimental comparisons with various models.

Keywords: dataset , word sense annotation , word sense disambiguation , network resource

* 通讯作者 Corresponding Author.
[1]https://github.com/SUDA-LA/Modern-Chinese-Word-Sense-Annotated-dataset

第二十二届中国计算语言学大会论文集，第43页-第53页，哈尔滨，中国，2023年8月3日至5日。

1 引言

词义消歧是自然语言处理中最经典的任务之一。它旨在识别多义词在给定上下文中的准确词义，以便更好地理解句子的含义(Weaver, 1952)。在汉语中多义词比例相对较低，但它们在自然语言中却被广泛使用。因此，词义消歧与语音识别、机器翻译、信息检索等领域密切相关，消歧的准确性直接影响这些领域的相关应用效率。

词义消歧数据集是词义消歧任务的基础语料，其质量关乎后续消歧任务的开展。鉴于其重要性，已有不少学者就如何构建高质量的词义消歧数据集展开了系统性的研究。一个完整的词义消歧数据集通常包含两种不同类型的数据集：1)词义数据集，包含待消歧词语的所有词义信息；2)标注语料数据集，能够将具有多种含义的词汇与其在语境中的正确词义联系起来，通常用于模型的训练和评估。

英文的词义消歧数据集建设已经趋于成熟，常用的词义数据集有WordNet(Miller et al., 1990)和BabelNet(Navigli and Ponzetto, 2012)。标注语料数据集则有许多不同的语料库可供使用，例如SemCor语料库(Miller et al., 1993)、Senseval-2语料库(Edmonds and Cotton, 2001)、Senseval-3语料库(Snyder and Palmer, 2004)，以及在语义测评比赛SemEval中使用的语料库：SemEval-2007 Task 17(Pradhan et al., 2007)、SemEval-2013 Task(Navigli et al., 2013)和SemEval-2015 Task 13(Moro and Navigli, 2015)，这些语料库都是基于WordNet中的词义构建而成。

相较于英文，汉语词义消歧研究起步较晚，数据资源相对匮乏。不过许多学者已开展相关工作并取得了不错的成果。例如北京大学的汉语词义标注语料库(STC)、汉语二语教学词义标注语料库、基于构词法的汉语词义消歧语料库(FiCLS)以及古汉语词义标注语料库。STC语料库(吴云芳and 俞士汶, 2006)使用《现代汉语语义词典》作为词义语料，对2000年1-3月和1998年1月的《人民日报》语料(共计642万字)进行多义词标注，标注了966个多义词义项。汉语二语教学词义标注语料库(王敬et al., 2017)和FiCLS语料库(Zheng et al., 2021)以《现代汉语词典》(CCD)为标注体系，前者对汉语二语教材文本(约350万字)中的1181个多义词进行标注；后者基于中文维基百科构建了有7064个多义词，囊括121655条标注数据的汉语词义消歧语料库。古汉语词义标注语料库(舒蕾et al., 2022)整合了多个词典资源，其语料库规模超过117.6万字，包含了3.87万条标注数据，极大的丰富了古代汉语领域的语言资源。

STC语料库构建时间较早，整体规模较小且缺乏时效性。汉语二语教学词义标注语料库和FiCLS语料库规模庞大，选用的都是近十年的数据作为标注语料，有较强的多样性和时效性，但二者都基于《现代汉语词典》进行语料标注，语料数据因词义数据集版权问题无法公开。因此当前汉语词义消歧任务面临高质量数据集获取困难的问题。同时本文注意到，鲜有数据集运用中文互联网资源，然而当前人们对一个词语的意思有疑惑时，会直接通过网络进行查询，从多个网络词典中找到最合适的解释。由此本文决定充分利用网络资源，基于网络数据构建一个高质量、公开的、由人工标注的且有一定规模的词义消歧数据集。

本文通过网络资源对多音节的多义词进行词义搜寻整合，筛选出1,083个词语作为待标注对象，从网络数据和专业语料中分别抽取句子开展词义标注，建成了超过85万字规模的现代汉语词义消歧数据集。以该库为基础，本文利用几种词义消歧模型进行词义消歧预测，其中模型F1值最高达到了77.74%。本文进一步分析了网络数据与专业语料的差别，证实了合理利用网络数据的可行性与可靠性。

2 网络词典爬取和融合

肖航和杨丽姣(2010)指出，构建词义消歧数据集存在两个难点。第一个难点是词典中的词义区分若不够清晰，可能导致标注结果不一致，第二个难点是词典提供的词义不够全面，会导致在标注时出现无法匹配的情况。同时，随着当前词义消歧技术的发展，义项作为词义的直观文字表示，其作用被逐渐挖掘。最近几年，向神经网络模型中添加义项信息来辅助消歧工作已成为主流，GlossBERT、BEM、ESCHER(Huang et al., 2019; Blevins and Zettlemoyer, 2020; Barba et al., 2021)这些充分利用义项信息的模型都取得了当时词义消歧任务的最佳水平。研究表明，当前神经网络模型能够有效的识别出不同的义项信息，同时义项信息对词义消歧研究有

项目资助：国家自然科学基金（62176173）、江苏高校优势学科建设工程资助项目

着极大的帮助。因此本文决定在构建词义数据集时，为同一词义添加多种文字描述，并尽可能提供全面的词义解释。

2.1 词义语料库选择

许多词典网站经过多年的更新和发展已经形成规模。本研究在经过深度分析及实际测试后发现，百度汉语[1]、在线新华字典[2]、汉典[3]这些在线词典词语丰富、词义质量较高且具备较高的专业性，但由于词义区分规则不同及构建人员文学认知水平存在差异，这些词典的义项大多不同，以“冲突”一词为例，百度汉语、在线新华字典、汉典分别给出了不同的义项，如表1所示。

网络词典	“冲突”义项
百度汉语	1.(动)矛盾激烈。
在线新华字典	1.有矛盾；争斗；争执。2.两种或几种动机同时存在又相互矛盾的心理状态。3.指文艺作品中人和人，人和环境，或人物内心的矛盾及其激化。
汉典	1.冲杀奔突。2.对立的、互不相容的力量或性质(如观念、利益、意志)的互相干扰。3.以争吵、摩擦和对立为特色的持久的不和。4.意见不合,发生争执。

Table 1: “冲突”在多种不同网络词典中的义项

汉典网站始建于2004年，是一个有着巨大容量的字、词、词组、成语等其它中文语言文字形式的免费在线辞典。其词典内的义项既参考了百度百科、维基百科等网络数据，又参考了王同亿编著的《高级汉语词典》，每一个词语义项都对有相应英文对应，英文解释是其进行义项区分的一个重要因素。汉典经过十多年的更新优化，其内部词义知识体系已趋渐完善，且义项充分，在使用者中获得了广泛好评。

不同于词义“粒度过细”的汉典，由无忧无虑中学语文网构建的在线新华字典词义更为凝练，针对同一个词义，在线新华字典与汉典的义项表述往往不同，正符合本文对词义数据集构建的需求。不过在线新华字典内部分词语存在义项冗杂、重复的问题，该词典在使用时需要提前进行专业的人工筛选。

在比对汉典与在线新华字典的数据特色后，本文拟将汉典与在线新华字典义项进行融合，充分利用不同词典的义项资源，在合并词典前后分别对两个词典数据进行人工清洗修改，以确保最终获得的词义数据集质量。

2.2 词语义项的获取及处理

2.2.1 网络词典的获取

为了获取汉典和在线新华字典的词典资源信息，本文根据两种网站的页面格式设计了不同的词语抽取策略。

在线汉语字典的网页源码采用HTML格式存储，整体风格较为简洁，包含了“基本解释”、“分解解释”、“相关词语”和“更多相关词语”等几个栏目。我们使用Beautiful Soup工具获取单个词语的网页源码，并通过正则表达式匹配获取“基本解释”中该词语的所有义项信息。在收集完当前词语的信息后，通过“相关词语”和“更多相关词语”中的链接依次爬取其他词语的网页源码，重复上述流程直到获得所有词语的信息。通过这种方式，本文共获得了368,024个词语解释，其中包含60,683个多义词。

汉典网站的词语解释页面功能较为丰富，单个词语的源码较大。为了提高爬取效率，我们采用了URL拼接的策略来获取特定词语的网页源码。例如，将“zdic.net/hans/”与“冲突”拼接即可组成汉典中“冲突”对应的URL：“zdic.net/hans/冲突”，使用Beautiful Soup工具便可获取“冲突”在汉典中的源码数据。采用这种方式对已经从在线汉语字典中获取的词语进行汉典源码爬取，取得源码中“词语解释”栏目中的相应词条作为对应词语的义项信息。基于这一策略，本文从汉典中共得到了302,273个词语解释，其中有62,620个词语是多义词。

[1]https://hanyu.baidu.com/
[2]http://xh.5156edu.com/
[3]https://www.zdic.net/

初步获得的在线新华字典和汉典数据存在诸多问题，包括网络词典本身存在的义项错误问题，以及由于爬取规则不全面引发的错误。为了使这些数据满足后续研究的需要，我们对各词典中多音节的多义词进行了系统整理。通过自动处理和专家人工审核相结合的方式对数据进行了清洗，确保各词典的词义满足以下要求：1)各个义项只包含词义的文本表示，不包含其他信息，如“英文解释”、“词性”、“音标”等。2)各义项表示的词义相互独立，不存在词义重叠的情况。3)如果义项中带有例句，则将各例句放在义项最后并用“|”分隔。最终得到的词典数据如表2所示。

网络词典	多义词数	词语词义数均值	单个义项长度均值
汉典	60,288	3.1	33.3
在线新华字典	58,995	2.4	36.2

Table 2: 处理后的网络词典数据

2.2.2 义项合并

在完成词典整理后，本文将对两个词典数据进行融合。一般的融合策略通常需要通过人工去重来保证融合后的语料中的每个词语义项都清晰可分。然而，在实际的融合过程中存在许多挑战。例如，不同词义的边界往往难以确定，同时两个词义可能存在交集，这给选择合适义项带来了困难。这些融合过程中的问题直接影响着后续的标注任务。因此，本文在融合词典数据时作出以下规定：1）允许在词义解释中存在表示同一词义的不同义项。2）不同词义可以存在交集。为了替代人工操作，本文采用了自动融合的方法，具体的融合流程如下所述：

步骤1：以在线新华字典数据为新词义数据集基础，将汉典中词语逐个添加至新词义数据集中，若原本在线新华字典中存在该词语词义解释，则跳转步骤2；若原本在线新华字典中不存在该词语词义解释，则跳转步骤3;

步骤2：将该词语在汉典里的义项逐个加入，加入时与在线新华字典中该词所有义项对比句子相似度，若句子相似度高于85%，说明二者表述极为相似，则不添加至新词义数据集;

步骤3：直接向新词义数据集添加该词语;

最终，本文得到了一个包含多种词义表述且有较广覆盖率的汉语多义词词义语料库，该语料库共有多义词59655个，每个词语平均有2.7个义项，每个义项平均有34.8个字，具体义项个数分布如图1所示。

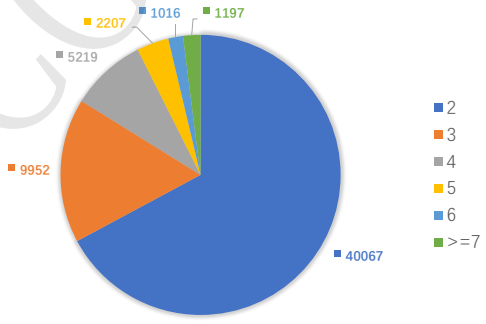

Figure 1: 不同义项数词语的分布情况

相比于人工融合，通过上述方法构建而成的词义语料库义项处理更为简单，极大的节约了词义语料库构建的人力物力成本。同时涵盖了更多更全面的义项，能够为当前深度学习模型的训练提供更多有价值的信息。

3 词义消歧数据集构建

3.1 待标注词语筛选

考虑到实际标注需要花费的人力成本，挑选1000个词语作为待标注词语，可以让每个词语都有足够多的标注语料与其对应，满足词义消歧模型的训练需要。同时为确保后续标注语料数据集的全面性和可靠性，本研究先对1998及2000年的《人民日报》中各词语进行词频统计，再将词语按照义项数分成六份(如图1中的划分方式)，依据词频抽取出这六份词语中各自出现最多的200个词语。

随后，本研究组织两名专家对这1,200个词语进行人工筛选，如果一个词语中只有一个词义是常用词义，其余词义在现代汉语中几乎不会被使用，则将该词语筛除。例如“等待”有两个义项：1.“不采取行动，直到期望或意料中的人、事务或情况出现；”2.“犹等到”。这里义项2出自《水浒传》第二回：“ 史进回到庄上，将陈达绑在庭心内柱上，等待一发拿了那两个贼首，一解官请赏。”，在现代汉语中“犹等到”这个义项几乎不再使用，若对含有“等待”的语句进行标注，标注结果单一，不符合词义消歧的需要，因此要将“等待”去除。最终，本文得到了1,083个包含多种常用词义的多义词作为后续标注语料数据集的主要标注词语。

3.2 标注语料采样及预处理

为确保标注数据的多样性和代表性，本研究拟从多种不同的语料库中抽取待标注语料。

《人民日报》和CoNLL2009中文语料是经过专业机构筛选和整理后的高质量语料库，有着规模大、内容丰富、文本质量高、语言风格统一的特点。而网络数据由本研究从各种网络资源爬取获得，是用户自由发布的文本内容，其包含大量的新闻报道、博客文章、舆论评价等信息，与另外两个语料相比，网络数据的语言风格更为多样化，其中词汇用法和上下文也更加复杂。采用多种语料库构建标注语料数据集，充分保证了数据的多样性。

虽然网络数据让标注语料更为多样，但其质量难以保证，存在错别字较多、句子过短、包含非文本相关的符号或标记等问题。针对这些问题本研究对获取的待标注句子进行人工预处理，确保即将被标注的句子满足以下原则：1)长度适中，待消歧词有较充足的上下文本。2)不含有错别字或逻辑错误。3)无与文本无关的内容。

基于上述原则，本研究从网络数据、1998年及2000年的《人民日报》语料和CoNLL2009中文语料中按照词语的五倍义项数抽取相应数量的句子，同时采用6：4的比例将网络数据和专业语料随机抽取的句子进行合并，得到24,455个句子。最后，对合并后的语料进行了人工处理，得到21,396个句子以供后续标注。

3.3 人工标注

本研究基于新构建的词义数据集和待标注语料，开展了语料标注工作。为确保标注质量，我们组织了共87名具有较强词汇理解能力的本科生和研究生参与人工标注。整个标注过程分为两个阶段。在第一阶段，我们使用了少量标注数据进行摸索尝试；依据第一阶段的标注经验，我们优化了标注流程和系统；第二阶段则展开了大规模的标注工作。

3.3.1 具体标注流程

参与标注的人员分为标注人员和审核人员，标注人员由82个本科生组成，审核人员由5个研究生组成。我们将对一个句子的标注称为”标注实例”，一个标注实例包含以下基础信息：1)含有待消歧词语的完整句子。2)待消歧词语的所有义项。每个标注实例会被分配给两个不同的标注人员，标注人员根据待消歧词语的上下文选择出适合此语境下的义项。如果两个标注人员选择的义项相同，该实例会被直接存储进数据集中。如果两个标注人员选择的义项不同，那该实例会进入审核流程，由审核人员进行判断并给出审核理由；经审核后的实例会再返回给标注人员，如果标注人员无异议，则进行学习；如果标注人员有异议，可以给出理由并进行投诉，被投诉的实例会交由另一个审核人员处理并给出最终答复。详细标注流程如图2所示。

3.3.2 标注规则

本研究构建的词义数据集包含重叠、相离和包含三种义项关系。重叠指的是两个义项的意思相同，但表述方式不同；相离指的是两个义项的意思不同；包含则指一个义项的意思被另一个义项所包含。以“接受”一词为例，其包含以下五个义项：1.收受。2.根据法令把机构、财产等拿过来。3.接纳。4.接纳；收受。5.依据法令收归己方所有。其中义项2和义项5之间存在重叠关

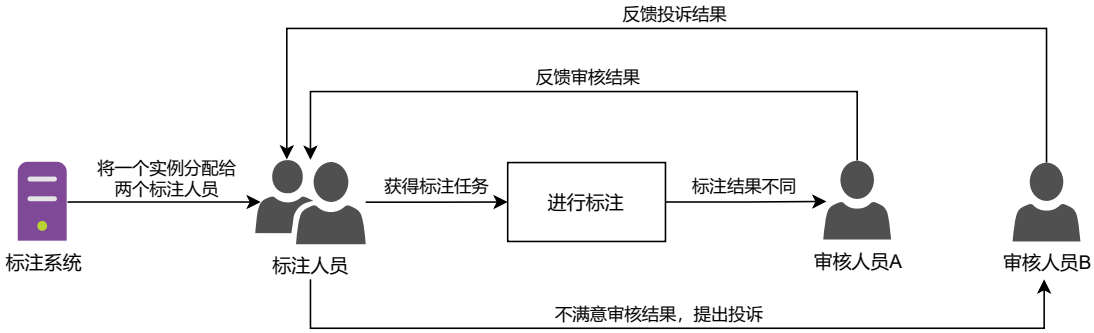

Figure 2: 标注流程

系，义项1和义项2之间是相离关系，义项1和义项4之间是包含关系。在标注过程中，标注者需要充分理解待标注句子，并根据义项的关系进行多选，具体多选规则如下：

规则一 不允许在一个标注实例中多选相离关系的义项。如果存在多个相离的义项都符合语境，则选择“待标注句子中词语有歧义”。

规则二 如果义项间存在重叠关系，则所有符合语境且重叠的义项都需要被选择。

规则三 如果义项间存在包含关系，则当被包含的义项被选择时，另一个义项也需要被选择。

规则四 如果只有一个义项符合语境，则单选该义项。

在实际标注中，有时候待标注句子可能存在一些本身就有错误或无法对应词义的情况。为了解决这些问题，我们特别添加了两个选项：“待标注句子本身有错误”和“没有合适的词义”供标注人员选择。同时，基于上述标注流程和多选规则，我们构建了一个轻量级的标注网站，使得人工标注变得更加方便和高效。图3展示了该标注网站的用户界面。

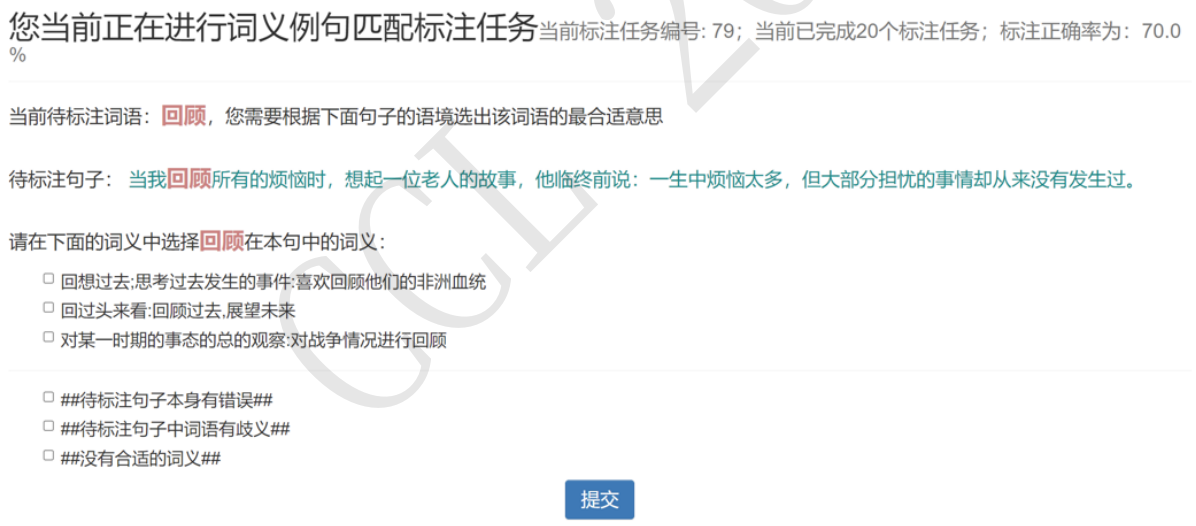

Figure 3: 标注系统界面

3.3.3 实际标注情况

在第一阶段的标注中，我们抽取了一小部分数据进行标注。每个标注人员需要逐一完成100个标注实例，这些实例是完全随机分配的。这一阶段的主要目的是筛选出准确率较高的标注人员，同时暴露出了以下问题：1)标注人员每次都需要学习新的词语义项，因此单个实例的学习成本过高。2)审核人员数量相对较少，标注人员完成标注后很难及时获得审核反馈。如果标注人员的词义理解出现错误，存在错误延续情况。

根据第一阶段的标注情况，我们挑选出了准确率较高的63名标注人员参加第二阶段的标

注。针对第一阶段中的问题，我们对标注系统进行了以下改进：1)将每个词语的所有实例分配给两个固定的标注人员，每个标注人员只有在完成一个词语的所有标注实例后才能开始下一个词语的标注。2)标注人员在获取一个词语的标注实例时，会先获得该词语30%的实例进行标注，只有在这些实例全部被审核完后，才会获得剩余的实例。在第二阶段，每个标注人员至少需要完成500个实例的标注工作。此外，在标注过程中，如果因为义项处理问题导致标注错误，这些错误也会被记录。在所有标注工作完成后，我们依据这些记录对词义数据集内的义项进行进一步优化处理。

所有标注完成后，我们对标注情况进行了统计。标注人员的平均标注准确率为74.6%，标注一致性达到了68.8%，进一步分析发现，一个词语前30%的实例的标注一致性仅为58.3%，标注者此时并未收到审核反馈，因此标注一致性较低，在学习审核结果后，标注一致性达到了76.2%。审核人员给出审核结果后，标注人员的投诉率为22.8%，其中有36%的投诉被采纳。从总体标注情况来看，标注人员对于有异议的审核结果能够积极反馈，且能从审核中充分学习到义项正确信息。

3.4 整体规模及义项分布

在本研究的标注过程中，有61名标注人员完成了词义消歧任务，共得到19,759个被正确标注的句子。其中，213个句子存在句子本身错误，160个句子中的词语具有歧义，304个句子无法找到合适的词义。最终，我们获得了19,082条高质量的标注语料，每条语料仅含有唯一的待消歧词，该词义消歧数据集涵盖了1,023个词语，总共包含4,831个义项，其中3,790个义项被正确标注的句子覆盖，义项标注覆盖率为78.45%。

数据集中各义项分布如图4所示，其中，有65.5%的义项在数据集中出现不足10次，且有18.4%的义项仅出现一次。对这些义项做进一步分析可发现：义项仅出现一次大多是因为该义项在近代几乎无人使用，但原词典包含了相应的例句，这些例句被收录在标注数据集中；一个词语的本义是最常被使用的词义，对应的义项在数据集中相较其它义项也更频繁。

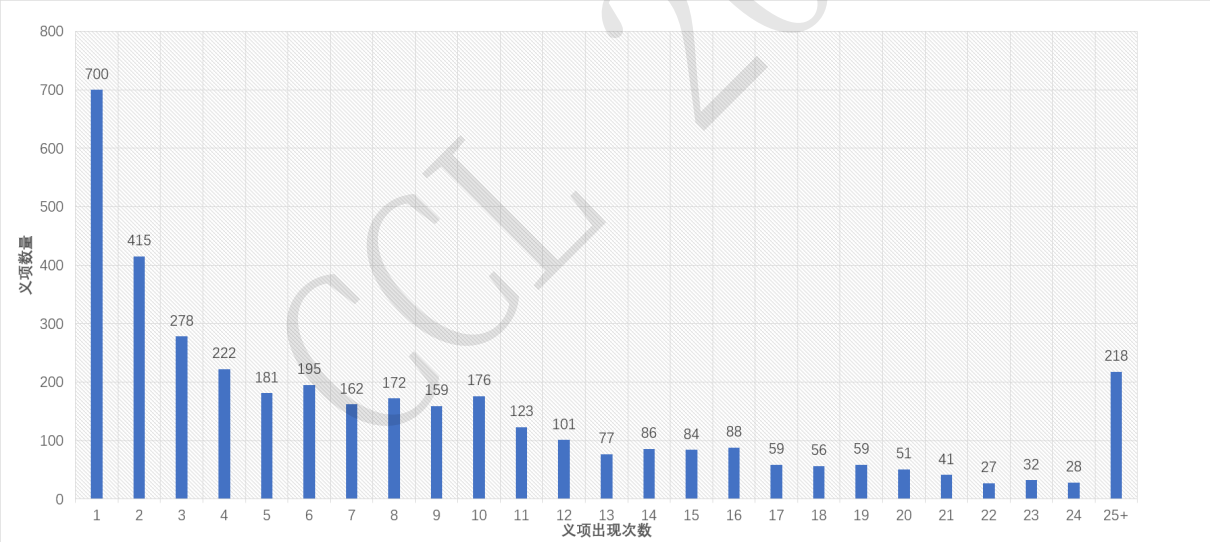

Figure 4: 数据集中各义项分布情况

3.5 网络数据与专业语料分析

本文构建的数据集共有19082条标注语料。其中，有9766条属于专业语料，10036条属于网络数据，这些语料的详细对比如表3所示。

可以看到，网络数据包含了更多的义项，极大的提高了整个数据集的词义覆盖率，同时，网络数据的格式与内容更为多样，弥补了专业语料格式单一的问题。在为数据集带来提升的同时，网络数据存在着以下问题：1)如图5所示，初始的网络数据存在很多噪音，需要花费较多人力进行清洗。2)网络数据的句长相对较短，句子提供的上下文信息有限。

语料资源	标注语句数	词语数	义项数	词义覆盖率	标注语句平均长度
专业语料	9,766	1,023	3,228	66.8	46.5
网络数据	10,036	1,021	3,689	76.3	41.3

Table 3: 网络数据与专业语料详细统计分析

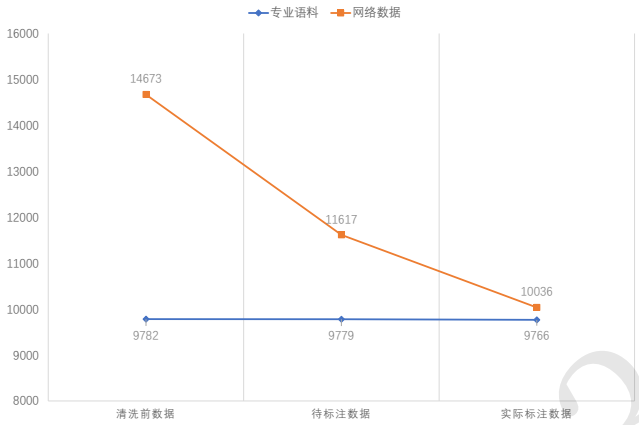

Figure 5: 标注数据规模变化

最终构建的标注数据集中，网络数据与专业语料占比近似1：1，这既保证了数据的专业性，又确保了语料的多样性和时效性。

4 词义消歧实验

4.1 模型介绍

GlossBERT 一种基于BERT(Devlin et al., 2018)的预训练模型，用于解决词义消歧问题。不同于传统的词义消歧模型，GlossBERT将义项信息也融入到神经网络中，并将词义消歧任务转化为句对分类问题。具体地，给定一个包含待消歧词w的文本c和其在词义数据集S中的对应义项$s_1, s_2, ..., s_n$，GlossBERT将c和S中各词义分别组合，并输入BERT进行句对分类。分类总共有两个标签：yes, no，分别表示s_n（是/不是）w的对应词义。该模型的创新之处在于首次将词语义项与BERT相融合，从而取得了较好的性能表现。

BEM 也是一种基于BERT的预训练模型。该模型采用双编码器的结构，将待消歧词w所在文本c和w对应的义项$s_1, s_2, ..., s_n$分别用BERT编码，并提取各自的[CLS]表示信息。相比于GlossBERT，BEM将所有义项都单独进行编码，有助于模型更好地理解不同义项表示的词义信息。实验结果表明，BEM在低频词上表现良好，进而提高了模型的整体性能表现。

ESCHER 是一种基于BART(Lewis et al., 2019)的预训练模型，它将词义消歧问题重新定义为一个跨度抽取问题。该模型把一个包含待消歧词w的文本c和w所对应的所有义项$s_1, s_2, ..., s_n$输入BART中，然后指定各个义项在模型输入中的位置信息，最终让模型生成与w最匹配的义项的位置。ESCHER首次将所有义项都放在同一个输入中让模型判断，相比于BEM，它在低频词上有着更好的效果。此外，它引入了高频噪音机制，通过向模型输入中随机添加高频词，从而变相降低高频词的选择率。在英文数据集上，ESCHER的F1值达到了80.7%。

4.2 数据集划分

本研究将之前构建的标注语料数据集按照7：1：2的比例划分为训练集，验证集，测试集。具体情况如表4所示。

数据集	词语数	义项数	标注语句数	标注语句平均长度	义项平均长度
训练集	918	3,259	15,244	44.2	11.0
验证集	604	1,453	2,893	43.1	11.7
测试集	678	1,778	3,645	44.6	11.6

Table 4: 数据集统计

为了评估模型在不同词义分布下的性能，我们采用了Zheng(2021)提出的测试集划分方法。将测试集按照词义在训练集中的出现频率分为四个子集：(1)最常用词义，即在训练集中对应词语出现最多的词义。(2)较常用词义，即在训练集中出现了超过5次，但不是对应词语出现最多的词义。(3)低频词义，即在训练集中出现过，但出现次数不多于5次的词义。(4)零样本词义，即该词义在训练集中从未出现过的词义。由于我们的测试集规模偏小，通过随机抽取的方式难以保证测试集中有足够数量的低频词义和零样本词义，因此本文在构建数据集时随机抽取了几个词语，将这些词语的所有语料放入测试集中，以此提高测试集中低频词义及零样本词义的比例。

值得注意的是，我们的标注数据集采用了多选策略，因此一个标注句子通常对应多个义项。同时，我们的数据集中待消歧词并不具有歧义，即每个标注句子对应的义项之间不存在相离的情况。这一特点保证了模型在训练过程中可以根据损失函数有效收敛。然而，这也需要我们在BEM 和ESCHER 模型加载数据时采用特殊的处理方式。

假定一个标注句子c的对应义项为$L = l_1, l_2, ..., l_k$。当使用BEM 或ESCHER 这类以最合适词义为输出的模型时，我们要对待消歧句子做单选衍变。一个标注句子进行单选衍变时会衍生为k个数据，这k个数据分别对应L中的一个义项，在进行词义预测时，对应词义数据会自动将L中其余义项从候选义项选择中去除。图6展示了这种单选衍变过程，图中原标注句子c有2个对应义项，4个候选义项。

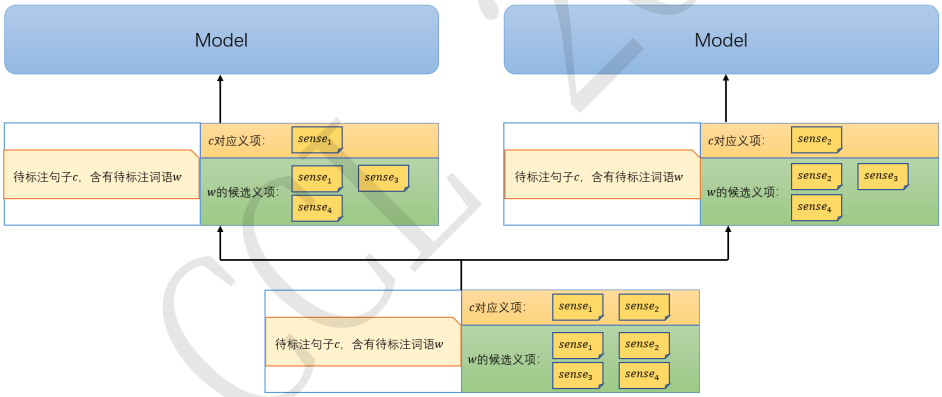

Figure 6: 数据单选衍变过程

4.3 实验结果分析

使用最常用词义(MFS)作为默认的基线参考，GlossBERT、BEM、ESCHER均采用原论文给定的参数设置，且都基于chinese-roberta-wwm-ext-large[1]进行训练微调。模型评估方面，我们以Macro-F1作为比较模型效果的依据。在本文构建的词义消歧数据集上，最终实验结果如表5所示。

相较于其它模型，GlossBERT的F1值偏低。在英文大规模数据集上GlossBERT有着不错的性能表现，但在我们的数据集上效果并不显著。从表5给出的结果不难发现，GlossBERT在低频词义及零样本词义方面效果较差，因此可以推断GlossBERT在小规模的数据集上难以学习到足够的词义知识。ESCHER是当前表现最好的模型，在小规模的数据集上，ESCHER已具备不

[1]https://huggingface.co/hfl/chinese-roberta-wwm-ext-large

模型	验证集	测试集	最常用词义	较常用词义	低频词义	零样本词义
MFS	26.78	41.85	×	×	×	×
GlossBERT	56.90	57.04	86.53	68.71	27.83	25.74
BEM	71.04	71.12	**92.32**	80.47	44.53	48.77
ESCHER	**76.81**	**77.74**	91.73	**84.38**	**52.57**	**54.53**

Table 5: 消歧模型结果

错的消歧能力，不论是BEM还是ESCHER，其模型性能的提升很大程度上源自于其在低资源词义数据上的改进。这种改进我们可以理解为是模型对于不同词义知识“理解”能力的提升。

考虑到标注语料的规模会对模型预测性能产生影响，我们从原1023个词语中分别抽取600、700、800、900个词语，依据4.2的划分规则对这些词语进行数据划分，并利用ESCHER模型分别训练、测试这些数据集，不同数据集规模的ESCHER表现如表6所示。

模型	总词语数	标注语料总数	测试集F1值
ESCHER_{600}	600	11,082	72.91
ESCHER_{700}	700	13,357	74.55
ESCHER_{800}	800	14,823	75.83
ESCHER_{900}	900	17,001	76.07
ESCHER_{all}	1,023	19,082	**76.81**

Table 6: 不同数据集规模下的ESCHER模型表现

当标注语料数不满15,000时，模型性能会随着语料的增加有较大提升，不过随着语料的不断增加，模型性能的提升幅度逐渐放缓，预估在标注语料数达到25,000时，模型性能受到语料规模的影响几乎可以忽略。不过值得注意的是模型性能在很大程度上也受限于数据集中的低频词义与零样本词义。因此，改善模型结构，提高模型对汉语词义的“理解”能力，让模型在更少量的数据上学习到更多的词义知识，是今后汉语词义消歧模型改进的一个重要方向。实践证明，模型能够从本文构建的现代汉语词义消歧数据集中学习到有效的词义知识，本文构建的数据集可以为后续的汉语词义消歧研究提供帮助。

5 结论

本文主要以现代汉语词义消歧数据集为研究对象，对两个公开的网络词典中多义词进行融合，筛选处理出1,083个词语作为待标注对象，以网络数据及专业语料作为源语料库进行标注语料抽取，并依据标注规则进行了人工标注。最终，本文构建的词义消歧数据集包含将近2万条标注数据，规模超过85万字。本文利用多种词义消歧模型对该数据集进行测试，既验证了数据集的质量，还探讨了汉语词义消歧模型发展的趋势。后续该数据集会无偿公开，方便更多学者进行汉语词义消歧研究。

不过本文构建的数据集规模依旧偏小，不太适用于对数据量有较大需求的模型，为了满足今后的科研需要，我们希望从以下几点继续改进现有的资源：1)对那些低频词义进行标注语料补充。2)用更多种类的语料扩充当前数据集，使语料覆盖更广。3)不局限于多音节词，对更多的多义词标注，努力实现全词标注。

参考文献

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021. Esc: Redesigning wsd with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss-informed biencoders. *arXiv preprint arXiv:2005.02590*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Philip Edmonds and Scott Cotton. 2001. Senseval-2: overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. Glossbert: Bert for word sense disambiguation with gloss knowledge. *arXiv preprint arXiv:1908.07245*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Andrea Moro and Roberto Navigli. 2015. Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 288–297.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 87–92.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43.

Warren Weaver. 1952. Translation. In *Proceedings of the Conference on Mechanical Translation*.

Hua Zheng, Lei Li, Damai Dai, Deli Chen, Tianyu Liu, Xu Sun, and Yang Liu. 2021. Leveraging word-formation knowledge for chinese word sense disambiguation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 918–923.

吴云芳and 俞士汶. 2006. 信息处理用词语义项区分的原则和方法. 语言文字应用, (2):126–133.

王敬, 杨丽姣, 蒋宏飞, 苏靖杰, and 付静玲. 2017. 汉语二语教学领域词义标注语料库的研究及构建. 中文信息学报, 31(1):221–229.

肖航and 杨丽姣. 2010. 基于词典的语料库词义标注研究. 语言文字应用, (2):135–141.

舒蕾, 郭懿鸾, 王慧萍, 张学涛, and 胡韧奋. 2022. 古汉语词义标注语料库的构建及应用研究. 中文信息学报, 36(5):21–30.