清华大学

博士学位论文选题报告

题 目：语言模型中词表征的选取和评估

院（系） 人文学院

学　科 中国语言文学

导　师 刘颖教授

研 究 生 刘柱

学　号 2022312212

报告日期 2024 年 7 月 1 日

目录

1 选题背景及研究意义

1.1 选题背景

当前的计算语言模型表现出了卓越且通用的语言理解能力。一方面，以 BERT[1]、GPT-4[2] 等为代表的模型在众多语义理解任务上取得不俗、甚至超越人类的性能，这些任务包括：词义消解[3]、机器翻译[4]、阅读理解[5] 等等。另一方面，这些模型在不同任务、不同语言、不同模型等中往往表现出强大的通用性和可迁移性[6]。由此，它们往往在人工智能的各个领域中作为基础模型（foundation model[7]），为之后的领域内相关任务提供初始化。

语言学家分层级研究语言的理解问题，这些层次包括词、短语、句子、篇章等等。其中，词义往往是语义理解中初始且重要的一部分。词是最小的能够独立活动的有意义的语言成分[8]，作为更高层级的基本单元参与到语义的组合性[9] 中。反过来，更高层级（例如短语、句子）提供的上下文往往影响同一个词汇形式的意义。这种依赖于上下文信息、形式与意义的“一对多”现象也被称之为词汇的“多义性”。它往往与心理词库的经济性原则相关[10]，即：重复利用同一个符号表达多样的意义，从而减轻人脑记忆的负担，也因此具有语言上的普遍性[11]。

语言学通过分析不同词汇实例[1]出现在上下文的位置和含义，区别词义，并按照意义之间联系的紧密程度归纳概括出不同的词义差异，它们由远及近可以分为：同形异义词（homonymy）、多义词（polysemy）、语义角色差异等。语言学家往往通过分析词汇语义的特征来判断这些差异，例如义素分析法[12]、原型特征分析法[13]、语义地图分析法[14] 等。这些分析方法具有上下文性、离散性和主观性的特点。

另一方面，当前先进的计算语言模型学习一个高维的实值向量来表征词汇。根据分布式语义假设[2]，这一随上下文变化的向量代表了词汇的上下文语义。它在某些维度上与语言学研究的特征分析法类似，例如：高维性和上下文性，但它是连续、客观且不透明的。同时由于当前语言模型的训练目标不同、架构不同、且具有分层性，反映词汇语义的向量如何提取仍然是一个未知的计算问题。

因此，探寻计算模型如何寻找表征，以及表征多大程度上可以反映上下文词义的差异，就成为理解模型的重要方面。这种探寻更大的背景是计算模型的可解释性问题。由于当前模型训练所需的庞大参数、海量数据、无需显性提取特征等特点，作为一个黑盒的模型学习到了哪些知识仍然存在很多疑问。可解释性研究试图从不同的角度理解机器的可知论问题。本文则着眼于语言学背景下的知识，具体是词汇的上下文语义知识。

[1]即形符（token-level），它与同一个词形的类符（type-level）相对。

[2]即出现在相似上下文的词汇具有相似的语义。[15-16]

1.2 研究意义

本研究作为计算机领域和语言学领域的交叉，通过分布式语义将二者结合在一起，并对两个方向都具有研究意义。

探究语言模型所掌握的语言学知识，有助于客观评估模型，从而增强模型的可解释性。当前大语言模型作为新一代人工智能的基础技术已经受到人们的普遍接受和青睐。然而，模型到底真正所掌握的知识有哪些？这些知识包括世界知识、常识知识以及语言学知识。本研究从语言学知识作为切口对大语言模型进行评估，并具体到词汇在上下文中的多义现象。通过研究不同范畴的多义现象，构建相应的数据集和评估标准，并参照不同的语言，来客观评估模型的性能，从而帮助研究者更加理性地认识模型，增加模型的可解释性，从而提高其可靠性和安全性。

通过语言模型可以进一步增强语言学中对于词汇多义性的理解。多义现象体现了形式和意义的不统一，是语言学家热衷研究的话题。语言学者通过显示分析同一词汇在不同上下文中的特征，辨析不同的语义，但语义的归纳难免存在主观性、不确定性。由于计算模型在大量语料上训练，因而可以为词义的辨析提供一个较为客观的参考。

2 文献综述

该章节介绍与论文相关的国内外研究现状，首先从语言学的视角分析了词汇多义性的几个方面，包括同源异义词、多义项词和语义角色。之后介绍了神经网络模型中可解释性的相关研究。最后概括了计算模型中对多义性的研究。

2.1 词汇多义性

词汇多义性是指同一个词汇单元，由于出现的上下文不同，而导致语义发生变化的现象。由于本研究侧重于意义的变化，并且为了保证跨语言的普适性，这里的“词汇单元”并不完全指可以自由独立使用的最小音义结合体（即，word），也可以包括在某些语言（如汉语）中不可以自由使用，但仍然具有实在意义的最小音义结合体（即语素）[3]。为了方便指陈，本文统一采用“词”或者“词汇”表示这一单元。

另一方面，本文研究的“意义”（meaning）不仅仅包括列入到词典中独立义项（sense）的那些较为明显的意义，也包括（原型）语义角色。后者往往不会全部单独在词典中列出，但仍在对比句对中发现它们的语义差异。

最后，本文讨论的词汇多义性往往受到上下文影响，这里的上下文常常指词汇所在的独立的句子，我们忽略更多范围的篇章上下文，以及言语会话中的情景上下文。同时，本文仅关注**同一个**词汇在不同上下文中的语义差异，不特别关注不同词汇之间的语义联

[3]由于缺乏显性标记，汉语对于词的鉴定相对困难，例如它和某些语素或者短语的区别。本文不专门讨论这一问题，对它的鉴定大多采用现成公认的工具书中的看法。

表 1 不同多义类型的区分

类型	语言平面	义项关联程度	是否需独立义项	研究单元
同形异义词	词	不相关	需要	词
多义项词	义项	相关	需要	词
语义角色	用例	紧密	不需要	主谓宾结构中的词

系，例如同义词、反义词、上下文词等，尽管后者也是词汇语义学中的重要部分。

根据词汇不同意义之间的相关性紧密程度，本文提炼出了四种多义现象。它们之间的紧密程度由弱到强，依次为同形异义词、多义项词和原型语义角色。后文依次对它们进行介绍，表 7列出了它们之间的对比。

2.1.1 同形异义词和多义项词

同形异义词（homonym），也即同音词[4]，是指词形和语音完全相同、但意义之前并无关联的词汇现象[17]。多义项词（polysemy）[5]则是指词形相同，意义不同但有关联的现象。一般来说，同形异义词的意义之间不具有相关性、不共享同一个词源，在词典中列入为独立的词条，从而被一般人认为是不同的词语。例如 bank 的银行义和河岸义；“花”的动词花钱义和名词植物义。而多义项词的词义之间关联很紧密，常常具有共享词源、引申、借喻等相似或相关的关系[12]。例如：position 既表示物理空间的位置，又表示职位上的位置；“头”既表示人体的身体部位，又表示领袖。早在古希腊时期，亚里士多德的《范畴论》中就提到了同名异义这一现象[18]，并用到了 homonymous[6]这样的术语。

根据 Lyons[19] 的研究，词典编撰者一般应用两条重要的原则来区分同形异义词和多义项词。一条是词汇的来源，一条是意义之间的相关性强度。然而，区分同形异义词和多义项词有时候并非十分容易。原因如下：一方面，很多时候共时上无法区别的义项，其实从历时来看是有关联的。例如文献[20] 中提到 bank 的银行义和河岸义在词源上具有相关性，但对于现在言语社团的成员来说，这种理据上的相关性已经模糊，从而把它认为是不同的词。另一方面语义之间的相关性判断具有主观性，有些词汇意义之间的相关性强度在不同成员之间的感受不同，例如汉语中的“死”的“失去生命”的意义和“不灵活”的意义之间的关联性就存在争议。

词类对于同形异义词和多义项词的区分也具有直接帮助，一般认为属于不同词类范畴的词属于同形异义词，尽管它们的语义关联度较大，例如英语中 talk 做动词讲是演讲

[4]英文的 homonym 是指词形相同，意义不同（例如 contract 的合同义和收缩义）或发音相同，意义不同（例如 sea 和 see）。由于本文不考虑语音，仅考虑文字上的差异，故特指前一种情况，即 homograph。

[5]一般将 polysemy 译为多义词，为了和本文提到的“多义性”区分，这里使用“多义项词”进行对译，突出了意义可以固定下来（established sense），而成为词典中的义项。

[6]原文中采用的是 homōnuma，并且仍有学者争论他在使用这个术语时候，实际所指的是同形异义词还是多义项词[18]。

义，但它同时具有名词的演讲义，但它们普遍仍归属于不同的词条[7]。这时候主要考虑到不同词类范畴的词在句法环境、形态变换等等形式上都具有明显的不同。在计算语义词典例如词网[21]中词类也作为重要的词汇特征，并在词义相关的计算任务上广泛应用。但是这一方法对于汉语来说仍存在较大争议[22]。汉语中存在大量的兼类词，例如大部分双音节动词都可以用作名词[23-24]，它们之间的意义对一般汉语母语者来说并没有显著差异。如果让它们视作同形异义词，也就是不同的词的话，那汉语的词库会增加很多没有必要的冗余。

同形异义词和多义项词的在某些情况下的区分困难关键在于词义划分以及词义相关性距离感知之间的主观性。本文采用典型范畴理论[25-26]，选择出典型的同形异义词和多义项词放置在连续变化的两端，同时一些非典型的案例则置于连续统的中间。同时即使都是多义项词，它们义项之间的相关性距离也存在差异，例如作为植物义的花义（花草）和作为人工品的花义（花环）的相关性要比具有条纹义的花（花边）更近。它们也处于词义变化连续谱的不同位置。

2.1.2 原型语义角色

一个词的语义也体现在句子中其他成分之间的关系中，这一关系可以通过语义角色进行刻画。语义角色（semantic role）描述了“主体对客体实施了什么行为，以及发生的时间、处所等条件”[27]。常见的语义角色包括施事、受事、时间、处所等。格语法（Case Grammar）以及语义角色由 Fillmore[28] 在上世纪 60 年代提出，提出用这些跨语言共性的角色来替代主宾语等句法概念。之后语义角色的概念在形式语义学派、以及认知语义学派都产生了不少影响，并进一步发展。

原型语义角色由 Dowty 提出[13]，认为对于基本的主谓宾句式，及物动词的两端存在原型施事、原型受事。并且该原型范畴是一个连续变化的量，其程度受到不同因素的影响。表2罗列了不同因素，不同因素的影响程度差异，决定了施事性、受事性以及（动词的）及物性强弱的变化。例如下述例子：

(a) 行人走 便道

(b) 行人踩 便道

(c) 行人砸 便道

从上到下动作对于受事的影响程度依次增加，其动词的及物性依次增强、对应的主语的施事性、宾语的受事性也在一次增强。此后，认知语义派开始更进一步分析施事和受事的语义和语用特征。施事方面，Jackendoff[29] 认为施事关乎行为者、意愿者和致使者三个概念。同时也有学者[30]从名词的角度，例如生命度、意志性等进行识别和预测施事，还有学者[31]从语用和句式的角度进行分析。受事并非是简单的施事特征的反面，即施事和受事的不对称性[31]。在 Dowty[13] 提出的一些特征维度，例如运动程度（位移还是静止），它们的语义存在一定对称，然而在“自主性”和“变化性”上，二者并不

[7] https://www.merriam-webster.com/dictionary/talk

表 2 原型施事和原型受事的典型特征

原型施事	原型受事
意愿性 (violation)	经历状态变化 (change of state)
感知性 (sentience)	递增性 (incremental)
使动性 (causation)	受动性 (causally affected)
移位性 (movement)	静态性 (stationary)
独立性 (independent existence)	依存性 (existence not independent of event)

完全对等。同时施事受事的语义离不开中间的动词的及物性（transitivity）的强弱。一些工作[32] 也从动词角度进行分析，例如动词的“自主性”。

不同语言对于原型语义角色的变化采用的手段不同，例如汉语缺乏形态的变化，一般仅通过语序区分原型施事和原型受事，例如主谓宾的基本语序情况下，典型的施受事分别处于前面的主语位置和宾语位置，这一配置对应于“无标记语序”[33]。同时施受事的程度变化往往没有词汇形态上的变化，一般和通过句法手段（例如增删标记词）来反映，例如把字句和被字句、双宾语句等[31]。对于形态丰富的语言，可能通过格标记来反映施事、受事、以及它们程度的变化[34]。

原型语义角色的强弱变化往往会引起一种“格配置变动”的现象[34]，即这种格配置与常规的施受事对应不同，例如“便道走行人”中作为施事的“行人”到了动词后面本来是受事的位置。这大多发生在处于语义角色程度连续变换的中间地带。例如常常处于“与事”角色（例如工具、地点等）的成分就因为处于强施事和强受事的中间态，而视情况使用施事格或者受事格[35]。汉语中的可逆句现象也属于其中一类。汉语可逆句[8,36] 是指主宾互置，而整体语义以及语义角色不发生变化的现象，例如“行人走便道”和“便道走行人”。此时逆置后的句子中的主语充当了受事、宾语充当了施事，与常规的“主语-施事”和“宾语-受事”的配置不同。可以构成这样互逆的谓词、主语、宾语往往都是非典型施事或者受事。

2.2 计算语言模型的知识评估

计算语言模型从语言数据中学习出语言的分布和模式，进而可以拟合现有的语言数据并对新数据进行预测。随着现有技术的发展，以大语言模型为代表的计算模型已经表现出出色的语言理解和表达能力，且可以以对话为基本模式胜任许多更多智能化的任务，例如情感分析[37]、意图理解[38]。也因此，以大语言模型为基座的人工智能技术正在以前所未有的速度向大众普及，因此非常要必要对计算语言模型中学习到的知识进行评估。进行全方位的评估的必要性列举如下：

（1）评估语言模型有助于帮助我们更好地理解计算语言模型的优点和缺点。人工智

能当前的主流技术深度学习就是在 ImageNet 这个大型测试集[39] 上首先得到验证，并由于极大降低以往方法的错误率，而得到这一领域人员的广泛关注。PromptBench[40] 这一评测基准表明大语言模型对对抗性样本较为敏感，提醒研究者设计更好的提示模版。

（2）一个科学、准确的评估基准也可以公平比较各个语言模型，从而促进模型的发展。自从 ChatGPT 作为大语言模型进入公众视野外，各类大模型竞相发展。为了更好比较彼此的性能，就需要一些统一的基准平台，例如 GLUE[41]、SuperGLUE[42]、MMLU[43] 等等。它们可以帮助我们更好地比较不同的语言模型。

(3) 良好评估语言模型也可以增加模型的解释性、鲁棒性和安全性。当前以深度学习为基础的计算语言模型由于庞大的参数空间而缺乏良好的可解释性，大语言模型时代拓展法则（Scaling law）驱动下的、在更大规模数据的训练过程加剧了模型的黑盒性，与知识驱动的透明建模的范式越来越远[44]。各个角度对语言模型的评估有助于发现模型的能力边界、增加模型的可解释性，从而使得人工智能系统更加鲁棒和安全。

评估模型主要涉及到评估的内容、评估的方法、评估的指标三个方面。接下来要依次介绍这些方面。

2.2.1 评估内容

评估模型的知识可以分为以下的类型：语言学知识、认知知识、伦理知识。

对**语言学知识**的处理和评估又被称为基础自然语言处理（foundemental natural language processing)[45]，它主要从如下的语言学平面进行分析。词汇方面，包括词形变化[46]、词类标注[47]、词义消歧[48] 等任务；短语和句子层面，涉及到句法解析[49]、主谓一致[50]、句对推理[51] 等；篇章层面则包括阅读理解[5]、长文本翻译[52] 等任务；语言类型学的角度涉及到语言共性挖掘等等[42]。由于语言学知识是模型评估最重要的一方面，早就从人工智能诞生伊始就有对词义消歧[53] 等的关注和评估。SemEval[54-58] 依托于国际比赛，是评估各个角度语言学知识的一个全面的测试标准。值得注意的是，现在的大语言模型技术已经体现了良好甚至类人的语言表现能力（language performance），单从语言学知识的角度，已经有研究证明可以通过图灵测试[59]。

认知知识关于模型的推理和泛化（generalization）能力。记忆（memorization）和泛化的平衡一直是机器学习一个核心的问题[60]。记忆是指模型记住训练集出现的样本模式，从而帮助在其未见过的测试样本上作出预测。然而过度的记忆则会导致过拟合 [8]，从而影响模型的推理能力。为了增强模型的推理和泛化能力，除了一些传统的方法，例如正则化[60]、Dropout[61]、层归一化[62]，还有在数据层面，数据增强[63]、设计思维链的提示语[64] 等方式。从任务层面，ALERT[65] 将推理能力细化为如下 10 个子任务：逻辑、因果、常识、蕴含、数学、溯因、空间、类比、论点、演绎，并分别设计不同的数据集。CURRICULUM[66] 则将语言学知识和认识统一在一起，并归纳了词汇、句法、语义、逻辑、分析、常识推理、理解 8 类，并都统一将它们视为自然语言推理任务（natural

[8] 训练集性能非常好，然后测试集性能非常差

language inference）。

伦理知识关注模型产生的内容是否符合人类的价值观（alignment），这主要是针对以生成内容为主的大语言模型的。RICE[67] 从四个维度进行分析这种对齐：鲁棒性、可解释性、可控性和伦理角度。伦理方面包括是否对特定群体展示出歧视[68-70]、对个体进行伤害和攻击[71-72]、缺乏多样性和平等等[73]，前人[74-75] 也做了更加详细的综述。大语言模型主要通过人类反馈的强化学习（RLHF，reinforcement learning from human feedback）的方式学习人类的价值。同时一些研究也通过改变模型的输出表征[76]、内部参数[77] 对模型的伦理相关的输出进行修正。

2.2.2 评估方法

对模型的评估方法包括行为主义方法、基于表征的方式和机械可解释性方法。

行为主义方法将模型视为黑箱，通过分析输入输出关系来理解模型。例如最小对立对分析[78]，敏感性和扰动分析[79] 来评估模型的鲁棒性和变量依赖性[80-82]。这种与模型无关的方法对于复杂或专有模型很实用，但缺乏对内部决策过程和因果关系的深刻理解[83]。

基于表征的方式利用模型中间输出的隐状态向量来评估模型的能力。可以使用无监督学习和有监督学习来两种方式。无监督学习不借助额外的模型和监督信号，直接对隐向量进行评估。例如 Word2Vec[84] 中利用隐向量的处于欧式空间的位置来显示它的类比能力（国王 + 男人-女人 = 王后），也有其他方式通过聚类[85]、独立成分分析[86]、差值向量[87] 来进行评估和分析。另外一种有监督学习也被称为探针（probing）的方式。它往往通过在一个额外任务上训练一个简单的探针模型，来进一步反映模型是否学习到这一任务。这些任务包括语义角色[88]、句法[89] 等任务，并且在 BERT 模型时代发掘了不少有趣的结论，例如认为 BERT 的层次结构对应了传统自然语言处理的各个流程[90]。基于探针的方式需要设计可控实验，来保证最终表现是模型原始表征的知识，而非从新的任务上学习而来的。

机械可解释性是一种自下而上的方法，试图从神经网络内部，例如神经元、特征、神经回路（circulate）来分析模型[91]。不同于前面两种方式，机械可解释性采用逆向工程的方法来识别完成特定功能的神经回路。它常常通过设立最小对立的两个输入，将想要研究的部分使用噪声或者其他单词替代，之后通过观察输出的差异来定位回路。这些回路可以对应于推理功能[92]、复制功能[93] 等等。同时这些方法还假设单一神经元具有叠加态（superposition）和多义性（polysemanticity），从而试图分离出不同的特征[94]。这类对模型的评估强调从模型的内部和结构出发，借鉴物理学、神经科学和系统生物学等跨学科领域，指导开发透明、价值对齐的人工智能系统[77]。

2.2.3 不确定性评估

在评估模型的时候，我们往往设计针对性的语言学任务，并重点评估模型在该任务上的准确性能，然而考虑到词汇在根据上下文赋予意义时候，无法十分确定义项的归

属，我们将这一性质命名为不确定性（Uncertainty）。在模型评估“准确性”的同时，理应同时考虑模型的不确定性。

从义项划分角度，词汇多义性中不确定性的来源有很多。其一是义项之间并非完全独立，不同义项之间的语义域可能具有重叠，甚至包含的情况。例如“帮忙”在汉语大词典[95]中提供了如下两个义项：

(a) 帮忙做事

(b) 在别人有困难时给予帮助

其中 (b) 义项就是 (a) 义项的一个下位概念。在常用的 WordNet[21] 词典资源中，有学者认为其义项划分过细，因此有工作将细粒度的义项合并到更加粗的粒度[96]。第二个来源在于有些义项之间本身就无法确定性的分割开，这样构成的义项集合也被称为“模糊集”（fuzzy set）[97]。由于义项体现的是人类对于概念的认知，而这一认知本身就可能是一个连续状态。比较典型的是语义更加空灵的虚词，受到语法化的影响，它们的语法意义往往相互关联，无法截然分开。

机器学习中同样关注不确定性，它是指模型预测结果中包含的不确定性，也即预测不确定性（Predictive Uncertainty）。它又可以被分为两种类型[98]：数据不确定性（data uncertainty）和模型不确定性（model uncertainty）[9]数据不确定性是指数据产生时固有的噪声，无法通过收集更多的数据来减少这种不确定性。二模型不确定性则是由于缺乏足够领域内合适的数据集，导致对数据认知有偏差，最终产生的不确定性，因而它是可以通过增加数据来减轻的。在回归任务中，不确定性可以使用方差进行表述，而在分类任务中使用异众比率（Variation Ratios）和预测熵（predictive entropy）。

主流的深度学习框架都没有将模型的参数空间视作随机变量，仅对数据的输出进行概率建模，容易得到未被良好校准的（ill-calibrated）的概率输出，从而其概率值不能真实反映不确定性[99]。研究者使用多种改进的方式估算不确定性。例如，使用贝叶斯神经网络对其参数进行概率建模[100-101]；利用 MC Dropout 方法去近似一个高斯过程，进而估计不确定性[99]；使用模型集成（model ensemble）的方式通过不同的模型的结果输出来估计[102]；大语言模型中设计不同的 Prompt[103]、计算结果句义的语义熵[104-105] 等。

本研究将词义本身存在的义项间的非独立性和模糊性视作机器学习研究中的数据不确定性，同时考虑训练过程中由于域偏移带来的模型不确定性。

2.3 计算语言模型中的词义消歧

词义消歧任务是指选择词汇所在上下文中的最恰当的语义，我们用它来检验模型是否可以有效处理词汇多义性的问题。这一任务尽管较为明确，仍有不同的研究侧重点，这与自然语言处理技术的发展、数据集资源的建设、研究者侧重的角度等都有关系。本

[9]数据不确定性又被称为偶然不确定性（aleatoric uncertainty），模型不确定性又被称为认知不确定性（epistemic uncertainty）。偶然和认知不确定性更能反映二者在不确定性来源上的本质差异。

节分别介绍用于词义消歧的常见语料库、知识库、不同分类角度下的词义消歧模型以及它们的表现性能。

2.3.1 语料库

语料库是指人类自然语言片段的集合，这里的片段和集合一般指句子和文本，它用来推断模型的参数，使得模型可以生成这些数据。如果模型训练的过程需要监督信号，文本需要提前标注好真实标签（ground-truth label）。在词义消歧任务中，需要确定待消歧的目标词，以及它在上下文中的意义，如果目标词为所有的实词（content words），即名词、动词、形容词和副词，这时称其为全词标注；如果是针对一部分特定的词进行标注，称其为部分词标注。表 3展示了相关的数量统计对比。

2.3.1.1 SemCor 英文版的 SemCor[106] 是由普林斯顿大学开发的、目前使用规模最大、最为普遍、最流行的带词义标注的平衡语料库。它是布朗语料库[107] 的一部分，涵盖了新闻、社论、小说等等的体裁，无论在数量和质量上都可以作为美式书写英语的一个代表[108]。这一语料库人工标注了所有的实词的词类和语义，涵盖了 226,040 条标注，共 352 篇布朗语料库的文章。其中，语义词典选择了同期的 Wordnet 1.4 词典[21]。

2.3.1.2 OMSTI OMSTI (One Million Sense-Tagged Instances)[109] 是基于 WordNet 3.0 标注的大规模语料库，它基于中英互译的句子集 MultiUN[110]，利用外部软件（GIZA++[111]）提取的英汉词汇对齐信息，半自动化地构建了一个语义标注的语料库。尽管这种方法可能会带来一些错误的标注，研究者表明在随机抽取的样本中，正确标注率可以达到 83.7%，同时它的规模更加庞大。

2.3.1.3 WNGC WNGC (WordNet Gloss Corpus)[10]将 Wordnet 中的样例和定义解释部分拼到一起，自动链接到 Wordnet 所对应的义项上面，从而形成的一个语义标注语料库。这一语料库主要通过自动的方式获取，它充分挖掘了 WordNet 中的词汇信息，不需要任何人工标注。由于在这一语料库训练不涉及人工标注和监督，因此往往被认为是从知识中进行无监督学习。这种方法收集的语料库也易于拓展到多语言上，这得益于多语言 WordNet 的开发和使用。

2.3.1.4 OntoNotes OntoNotes 5.0[11]是由 BBN 科技、科罗拉多大学、宾夕法尼亚大学和南加州大学信息科学研究所共同开发的语料库。它覆盖了多样体裁的文本，包括新闻、手机对话、博客、新闻博客和脱口秀，支持英文、中文和阿拉伯文三种语言，标注了部分词汇的结构信息（包含句法结构和论元结构）以及语义信息。其中英文部分（大约 150 万英文单词）的语义标注采用的词典是粗略词义的 WordNet。

[10] https://wordnetcode.princeton.edu/glosstag.shtml

[11] https://catalog.ldc.upenn.edu/LDC2013T19

表 3 不同语料库的数量统计对比

语料库	文章	语料句子	语料单词	标注	词义种类	词种类[12]	多义度
Senseval-2[54]	3	242	5,766	2,282	1,335	1,093	5.4
Senseval-3[55]	3	352	5,541	1,850	1,167	977	6.8
SemEval-07[56]	3	135	3,201	455	375	330	8.5
SemEval-13[57]	13	306	8,391	1,644	827	751	4.9
SemEval-15[58]	4	138	2,604	1,022	659	512	5.5
SemCor[106]	352	37,176	802,443	226,036	33,362	22,436	6.8
OMSTI[109]	-	813,798	30,441,386	911,134	3,730	1,149	8.9
WNGC[113]	-	-	1,621,000	449,000	-	-	-
OntoNotes[113]	-	-	1,500,000	-	-	-	-

表 4 不同类型的知识库的数量统计对比

知识库	词条数目	同义词集合数量	单义词数量	多义度[13]
WordNet 3.0[14]	155,287	117,659	101,863	2.50
HowNet[15]	237,974 [16]	35,202 [17]	-	-
SyntagNet	-	71,025	-	-
BabelNet[114]	-	22,130,060	-	-
BabelNet_EN	-	13,964,713	-	-

2.3.1.5 SemEval 语义评估测试集（SemEval）是在过往的语义评估竞赛中设计的，成为了测试语义消歧任务的公共测试集。有相关工作[112] 对它们进行了整理。它一般由五次比赛构成，分别为：（1）Senseval-2[54]；（2）Senseval-3[55] 的任务 1；（3）SemEval-07[56] 的任务 17；（4）SemEval-13[57] 的任务 12；（5）SemEval-15[58] 的任务 13。

2.3.2 知识库

不同于非结构的文本语料库，知识库往往构建不同实体以及它们之间的关系，因此往往是一个结构化了的图结构，以下分别介绍了常用到的三种知识库，包括 WordNet、知网和 BabelNet，相关统计量分析可以参照表 4。

[12]将每个出现的待标注词如果它们具有相同的原型，就归入一类词。

[13]多义度是指平均一个词条包含的可能义项的数量，这里排除掉单义词。

[14]https://wordnet.princeton.edu/documentation/wnstats7wn

[15]https://openhownet.thunlp.org/

[16]包含中英文

[17]HowNet 并没有同义词集合的概念，这里指它总共的概念数量。另外，知网中包含 2,540 个义原。

2.3.2.1 WordNet WordNet[21] 是由普林斯顿大学开发的一个大型结构化知识词典。与普通的学习者词典不同，它是以同义词集合（Synset）为点、集合间的关系为边，具有图结构的词典。其中同义词集合代表一个概念，它们拥有相同的义项，通常用一句简短的语句以及少样的示例进行描述[18]。这些描述更多体现了非结构化的语言学知识（Linguistic Knowledge）。同义词集合之间又有不同的语义关系，WordNet 列出了如下的关系：（1）上下位关系；（2）部分-整体关系；（3）相反关系（仅针对形容词）。这些关系由于来自对于所指关系的认知，往往体现了结构化了的世界知识（World Knowledge）。

由于 WordNet 的组织格式更加适用于计算处理，研究人员已经开发了超过 60 种语言的 WordNet[19]。其中中文词网包括由南洋理工大学计算语言学实验室开发的中文开放词网 CoW[20]和台湾大学语言所研发的中文词网[21]。

2.3.2.2 知网 知网（HowNet）[22] 最早是由董振东和董强先生在 20 世纪 90 年代设计和构建的一部更加适用于中文的语言知识库，它利用常见汉字构建出最小的语义单元（即义原）集合，并利用它们对十几万的中英文词条进行语义标注。知网作为一个大型知识库，体现在一方面，义原标注采用了较为结构化的方式来进行，即罗列属性和对应的属性值（或也称为特征）以及复杂的语义角色关系的方式。[22]另一方面，2500 多个义原概念之间也存在多种关系，例如上下位关系、同义关系、反义关系、对义关系等。

2.3.2.3 BabelNet BabelNet[114] 是由罗马第一大学团队开发的、目前规模最大的、覆盖语言最广的知识库。它以英语 WordNet 为基础，在原先的英语同义词集合中融入更多异质的语言和百科资源，包括维基百科、维基数据、维基词典等，利用多语言 WordNet、百科的多语表达和自动翻译技术，覆盖了多达 520 种自然语言。相比 WordNet 仅利用词汇知识，BabelNet 有效利用了世界知识，这包含对概念或者命名实体的知识性描述以及多模态资源（例如概念对应的图片）的利用，从而有助于建设更加通用的、密集的知识网。

2.3.3 模型方法

词义消歧按照所利用的主要资源，可以分为监督式、半监督或者无监督任务、完全知识驱动的任务。其中完全知识驱动的方法不依赖于训练语料，仅通过探索大型的知识库来找到词汇-语义映射。监督式方法依赖带有词义标注的语料库，通过学习到一个词汇到语义映射的模型来解决这个问题。根据其具体任务的定义不同，又可以分为分类任务、语义检索任务、截取式任务和生成式任务。注意，这里的监督式方法不完全是单纯

[18]同义词之间仍享有完全一样的示例，话语解释和示例共同被称为 gloss.

[19]http://globalwordnet.org/resources/wordnets-in-the-world/

[20]https://bond-lab.github.io/cow/

[21]https://lope.linguistics.ntu.edu.tw/cwn/

[22]不同于 WordNet 按照描述性短语的方式释义，知网的释义方式更在于区分相同词语的不同的意义，而非准确地描绘出来。

表 5 词义消歧的常见方法总结。其中 SC 表示 SemCor 语料库，G 表示 WNGC 语料库，WN 代表 WordNet。知识资源中列举了三类，包含定义、用例和关系。

任务类型	训练过程	方法	训练语料	词典	知识资源		
					定义	用例	关系
监督式分类任务	分类任务	GAS[115]	SC	WN	✓	✗	✓
		GlossBERT[116]	SC	WN	✓	✗	✗
		EWISE[117]	SC	WN	✓	✗	✓
		EWISER[118]	SC+G	WN	✓	✓	✓
		MLWSD[119]	SC	WN	✗	✗	✓
		MLWSD*	SC	WN	✓	✓	✓
		RTWE[120]	SC	WN	✓	✗	✗
		RTWE*	SC+G	WN	✓	✓	✗
	语义检索任务	BEM[121]	SC	WN	✓	✗	✗
		Z-reweight[122]	SC	WN	✓	✗	✓
		SACE[123]	SC	WN	✓	✗	✓
		SACE*	SC+G	WN	✓	✗	✓
		ARES[124]	SC+WK	WN	✓	✗	
	截取式任务	ESCHER[125]	SC	WN		✗	✗
		ConSec[126]	SC	WN	✓	✗	✗
		ConSec*	SC+G	WN	✓	✓	✗
		KELESC[127]	SC	WN	✓	✓	✓
	生成式任务	Vec2Gloss[128]	-	-	✓	✗	✗
		Generationary[129]	CHA+SEM	-	✓	✗	✗
半/无监督	-	WSD_TM[130]	WK	-	✓	✓	✓
		WSD_LSA[131]	SE10	-	✗	✗	✗
完全知识驱动	基于相似性	Lesk[132]	-	WN	✓	✓	✗
		Lesk_ext[133]	-	WN	✓	✓	✓
		SREF[134]	-	WN	✓	✓	✓
	基于图算法	UKB[135]	-	WN+ESWN+EXWN	✓	✓	✓
		Babelfy[136]	-	BN	✗	✗	✓
		SyntagRank[137]	-	WN	✓	✓	✓
		WSDG[138]	-	WN+BN	✓	✗	✓

的数据驱动，融合了知识库的方法也归入到这一类中。半监督或者无监督式方法无需语义标注，它们仅仅从大规模数据中学习词汇的分布，就可以学习如何选择一个最佳的语义。

2.3.3.1 完全知识驱动 完全知识驱动的模型无需标注的语料库，仅通过探索知识库来推断词汇在上下文中的语义，常见的知识库格式与 WordNet 相关，包含词汇的定义（常常用人类理解的短文本来表示）、包含该词汇的用例、以及概念之间的关系，不同的模型使用不同方面的知识，表 5做出了对比说明。从算法的角度，本文将这类方法分为基于相似性匹配和探索图模型算法两大类。

基于相似性匹配的方法。早期的知识驱动算法基于语义连贯性假设：只有当一个句子所有词汇都被正确地消除歧义了，整个句子语义才是连贯的，进而每个词汇才算消除了歧义。这类似于一个词义标签序列的结构化输出问题。因此这些算法考虑句子中所有实词的任意的词汇对 $< w_i, w_j >$，并分别计算出这两个词汇所有可能组合的语义的相似性度量评分 *score*，选择评分最大的这组语义组合分别作为这词汇的最佳语义。以下为测度 $Score$ 的定义：

$$score : \mathcal{Z} \times \mathcal{Z} \rightarrow [0, 1]. \tag{1}$$

不同的方法考虑不同的相似性方法。Rada 等人[139] 以及 Leacock 和 Chodorow[140] 的工作将 WordNet 中概念上下位图的最短距离用作相似性度量的指标；Lesk 方法[132] 则计算由定义和示例组合在一起的句子间的重合程度，二者重合度越大表明词义越接近。拓展后的 Lesk[133] 也会融入知识图中的知识信息。这些基于语义连贯性假设的做法都需要同时考虑所有词汇的可能词义，对于一个长为 N 的句子而言，假设每个词汇大约有 k 个词义选择，那么模型处理单个句子的复杂度为 $O(k^N)$，这对于长句子而言，复杂度不可容忍。为了解决这个问题，一些方法[48] 不再显式考虑不同词汇之间语义依赖性，而直接使用该词汇的上下文和该词汇的可能语义做相似性计算，从而将时间复杂度降到 $O(N \times k)$。SREF[134] 利用 BERT[1] 提取上下文和语义（由定义、示例和一句相关的句子拼接而成）的表示，性能比较突出。

基于图算法的方法。由于知识库往往可以看作节点是概念，边代表关系的一个图，很多与图相关的算法也可以被直接利用。UKB[135] 算法采用随机游走的方式，并利用个性化 PageRank 算法来得到候选语义的排名；Babelfy[136] 则将词义消歧任务和命名实体识别任务结合，将仅仅包含语言知识的 WordNet 拓展到了带有百科知识的 BabelNet 上面，利用群团近似（clique approximation）的方法来学习知识图的信息；SyntagRank[137] 在原有的范式关系（paradigmatic relations）中添加了组合关系（syntagmatic relations），即更多考察了词语周围上下文的关系，为了获取这种关系，它使用了一个概念搭配相关的网络，即 SyntagNet[141]。它采用的图算法仍然是个性化 PageRank 算法。WSDG[138] 利用博弈论将词义选择过程视作一个博弈过程，并有效利用了 WordNet 和 BabelNet 中的关系信息。

融合语言学知识的方法。一种常见的与词义相关的语言学知识是词义搭配的选择偏好（selectional preference）或者选择限制（selectional restriction），即与某个词搭配使用时，之后的语义更偏向的范围，或者限制的范围。例如：“吃”只能与可食用的事物搭

配，这样与之不匹配的那些语义类便可以排除掉了。通过大规模语料可以使用经验频率去逼近存在某个语法依赖（例如动宾关系）的词与词之间的出现概率[142]，之后再通过词到语义类的映射，可以从候选的语义选项中选择一个最佳的语义。关于词到语义类的映射，可以采用多种方式，包括使用最小描述距离[143]，隐马尔可夫模型[144]，基于类别的概率[145] 和贝叶斯网络[146]。然后这类方法的效果被证明不如其他类型的基于知识的方法。

2.3.3.2 监督式数据驱动算法 监督式数据驱动算法依赖于研究者对于词义消歧的具体定义，不同的定义往往会产生不同的监督信号，从而使它们在数据格式、模型设计以及损失函数的具体实现上都有差异。常见的算法大多采用**分类**的任务，即通过带参数 θ 的模型，计算出定义在整个语义选项空间 $\mathcal{Z}$ 的一个多项分布，该分布表示模型选择对应语义的概率：

$$p_\theta(\mathcal{Z}|c, w, \mathcal{E}). \tag{2}$$

这一任务往往采用分类的交叉熵损失。另一类算法将其定义为一个**语义检索任务**，即从一个待选集合中检索出与上下文语义最接近的一个语义选项。这类算法往往需要训练一个上下文表示模型 q_α 和语义表示模型 k_β，将前者输出的表示用于查询，后者输出的表示用于匹配，匹配的过程采用 1-近邻算法即可：

$$1nn(q_\alpha(w, c, \mathcal{E}), k_\beta(\mathcal{Z})). \tag{3}$$

该类任务可以采用检索常用到的例如对比损失、最大间隔损失（max-margin loss）等。受到其他自然语言任务（例如文本问答）的影响，有些算法采用**截取式任务**定义，即将所有候选定义拼接在一起，目标是得出目标定义的索引位置。另外也有算法采用**生成式任务定义**，即目标是生成一个语义定义。

分类任务 与人工智能的发展一致，分类模型大致经历了三个时期：早期的规则设计时代，中期则基于特征选择和模式识别，近期随着训练数据的激增和算力的增强，进入了深度学习时期。词义消歧作为一个古老的人工智能任务[53]，其解决方法也历经这些时期。规则设计时期主要是设计语言学规则，例如词类、词的语法功能等来推断词义[147]。规则设计的繁多和语言的复杂性等因素导致这类方法无法进一步拓展，伴随着人工智能的发展也陷入低谷。统计时期得益于较大规模的语料，例如 SemCor[106]，很多机器学习算法得以应用于这些数据的模式挖掘上。相关算法包括决策树[148]、朴素贝叶斯模型[149]、（浅层）神经网络[150]、k 近邻算法[151]、SVM 算法[152] 等。这些方法仅仅用到语义标注的语料库，这基于语义分布相似性的假设，认为词汇语义仅从上下文中可以学习到。它们普遍存在知识瓶颈问题，即获取大量标注的数据的困难性；以及无法解决语义分布不均衡的问题。

分类模型在深度学习时期主要采用深度神经网络进行分类，同时很多分类模型开始探索将知识库的信息融合到以往仅依赖语料的模型中，这些信息包括定义语句信息、

示例信息和关系信息（参见章节2.3.2.1WordNet 的介绍）。GAS[115] 利用长短时记忆模型（LSTM），将定义信息通过记忆模块融入到分类模型中。GlossBert[116] 将定义信息拼到上下文中，将任务定义为语义和词汇是否匹配的一个二分类任务。之后的模型大多利用预训练语言模型挖掘更多上下文之间的关系。EWISE[117] 编码了定义的语义向量，并把这个信息融入到分类模型的输出中；它的改进版本 EWISER[118] 进一步利用知识库中的关系信息，融入更多相关的定义向量，同时还把定义和用例拼接到输入数据中。MLWSD[119] 则观察到不少的样例的标注不止一个正确标签，因此将这一任务定义为多标签的分类任务，同时它也利用知识库中的关系信息找到了更多相关的标签。RTWE[120] 利用定义与目标词汇之间的相关性，应用迭代式的注意力机制将定义的信息融合到模型中间的输出层中。

语义检索任务 基于语义检索相似性可以充分利用知识库中的定义信息，从而缓解稀有词义分布不均衡的问题。BEM[121] 利用两个编码器分别编码语义句子和上下文；之后有工作[122] 基于 BEM，利用重采样的方式显式地缓解语义分布不均匀的问题。SACE[123] 发掘上下文词汇的词义间的依赖性，并利用句子的相似性，找到更多的上下文。ARES[124] 强调上下文的组合关系也同样重要，并从网络（例如维基百科）上检索到更多相关的上下文，以得到更加丰富的上下文特征向量。

截取式任务 截取式任务除了可以利用正确语义的句子信息，还可以将所有候选的语义句子信息利用上，从而可以更加缓解稀有语义分布较少的问题。ESCHER[125] 首次将截取式任务应用到词义消歧中，有效解决了词义分布不均衡的问题。它的改进版本 ConSec[126] 则更多地利用了上下文中已经消歧了词汇语义信息，从而确保了语义的连贯性。KELESC[127] 在 ESCHER 的基础上，从模型的输入部分更多地融合了通过知识库的上位关系检索出来的其他词义信息，从而使模型看到更多上下文。

生成式任务 生成式任务通过建模词汇序列，从一个上下文向量表示中直接生成定义，这也被称为定义建模问题（Definition Modeling）。早期的生成任务[153] 通过设计词汇语义规则模版，生成特定属性的语义表示。近期的生成任务的定义为人类可读的句子，这一任务一开始主要是为了静态词向量的可解释性的[154]，之后也被用于动态的上下文向量中。针对于词义消歧任务的定义建模则往往定义一个“编码-解码”模型[128]，即输入输出都是一个词汇序列。这种序列生成模型可以由早期的循环神经网络[154]，长短时注意力机制[155]，到现在基于自注意力机制的 Transformer[156] 来实现。近期的做法[129] 将作为条件的嵌入改为一个目标词的起止索引，用来引导之后的生成。多义词的语义生成需要考虑到词的多个义项的分布，这些分布可以通过一个离散[155] 或者连续[157] 的隐变量生成模型去学习，或者用近似的分布[158] 去逼近一个混合高斯模型。根据不同语言的特性，这一任务也可以在跨语言或者其他语言中应用[159-161]。

2.3.3.3 无/半监督式数据驱动算法 无监督算法无需词义标注，不存在知识瓶颈的问题，它也不需要提前已知的候选词义集合。这类算法通过发掘语料库中相似的上下文，自动得到不同的语义组，每一个语义组代表一个语义。它是词义消歧任务的另一种代表形式，也被称为词义归约。早期的算法选择较为简单的上下文表示，例如 N 元组的词频等，后期可以选择更加复杂的表示，包括静态表示如 Word2Vec[84]，Glove[162]；上下文表示，例如 Bert[1]。规约的算法可以选择聚类算法[85]，或者隐变量模型[130-131]。

半监督算法只需少量的数据标注，通过迭代的方式生成更多标注的数据。常见的算法有 bootstrapping，它包括联合训练[163] 和自训练[164] 两种方式。也有方法利用不确定性采样和主动学习[165] 的方式，挑选信息量丰富的样本来标注。外部数据增强的方式也可以看作是一种半监督的方式，例如借助成对的语料库，迁移学习[166]，或者大语言生成模型[167]，例如 GPT-2 等方式。

2.3.4 性能比较

词义消歧模型一般在一个统一、公有的数据集或者评价基准上进行比较，这些数据集主要来自过往语义评估测试集 SemEval（参见2.3.1.5）。为了统一标准和简化测试流程，研究者[112] 开发了一个统一的评测平台：近期的主流方法都在这个平台上进行测试。测试集上的测试一般都可以视作是一个分类任务，因此采用分类任务常用到的 F1 评分[24]进行评测。表 6列出了常见方法对应的 F1 评分，包含各个常见的测试集，以及按照实词的词类进行分类后的结果。除了上述常规的方法，文章也考虑以下几种常见的基准模型：

上界水平 这一任务的上界水平参考人类标注者的表现，词义标注一般由多个人标注，只有都所有人都达成一致时，才会有把握地将其标注为这个一致的义项。由于词义的模糊性、词义的连续性或者自然语言固有的歧义，导致有些词汇的词义判断困难，相比一般的分类任务，有更大地比例无法达成一致，而一般倾向于认为模型无法超出人类达成一致的比例，或是即使超出，也无法解释。所以采用标注者相互一致性（ITA，inter-tagger agreement）作为模型的上界水平，据估计，这一水平大概在 80% 左右。

下界水平 下界水考虑模型无需训练就可以计算出的结果，本文采用测试集中的单义词的比例(由于单义词仅仅有一个义项,故对它的选择一定是正确的),记做 LB_Mono。

占优基准 占优基准是指考虑到无需训练，仅凭借一定的先验知识就能获得的具有竞争力的方法。这包括直接选择最常见义项（MFS, most frequent sense）。最常见义项可以通过训练库中的统计（MFS_Cop）或者通过 WordNet 的第一个义项（MFS_WN1）得知[25]。还有一个方法是 ChatGPT[26]，它通过网上在超大规模语料进行预训练学习到的通用性人工智能模型，具有零样本学习的能力。

[24] 多数方法仅仅包括其 micro F1 评分，它是从词汇的角度进行的评测，另一类 macro F1 评分则是从词义类别的角度进行的评估，更加可以加重对于不平衡语义选项的识别[168]。本文仅仅考虑 micro F1 评分。

[25] 根据 WordNet 的编纂特点，第一个义项是根据词频决定的。

[26] http://chat.openai.com

表 6 各类消歧方法的性能对比

方法	SE02	SE03	SE07	SE13	SE15	ALL	名	动	形	副
ITA[118]	-	-	-	-	-	80.0				
LB_Mono[23]	-	-	-	-	-	17.4	13.5	4.5	23.7	16.3
MFS_Cop	65.6	66.0	54.5	63.8	67.1	65.5	-	-	-	-
MFS_WN1	66.8	66.2	55.2	63.0	67.8	65.2	-	-	-	-
ChatGPT	-	-	-	-	-	73.3	-	-	-	-
GAS[115]	72.2	70.5	-	67.2	72.6	70.6	72.2	57.7	76.6	85.0
GlossBERT[116]	77.7	75.2	72.5	76.1	80.4	77.0	79.8	67.1	79.6	87.4
EWISE[117]	73.8	71.1	67.3	69.4	74.5	71.8	74.0	60.2	78.0	82.1
EWISER[118]	80.8	79.0	75.2	80.7	81.8	80.1	82.9	69.4	83.6	87.3
MLWSD[119]	78.4	77.8	72.2	76.7	78.2	77.6	80.1	67.0	80.5	86.2
MLWSD*	80.4	77.8	76.2	81.8	83.3	80.2	82.9	70.3	83.4	85.5
RTWE[120]	83.4	82.9	74.5	82.1	85.3	82.7	84.9	72.8	87.7	87.9
RTWE*	85.2	83.3	77.1	83.8	86.3	84.1	85.7	75.1	90.6	88.7
BEM[121]	79.4	77.4	74.5	79.7	81.7	79.0	81.4	68.5	83.0	87.9
Z-reweight[122]	79.6	76.5	71.9	78.9	82.5	78.6	-	-	-	-
SACE[123]	82.4	81.1	76.3	82.5	83.7	81.9	84.1	72.2	86.4	89.0
SACE*	83.6	81.4	77.8	82.4	87.3	82.9	85.3	74.2	85.9	87.3
ARES[124]	78.0	77.1	71.0	77.3	83.2	77.9	80.6	68.3	80.5	83.5
ESCHER[125]	81.7	77.8	76.3	82.2	83.2	80.7	83.9	69.3	83.8	86.7
ConSec[126]	82.3	79.9	77.4	83.2	85.2	82.0	85.4	70.8	84.0	87.3
ConSec*	82.7	81.0	78.5	85.2	87.5	83.2	86.4	72.4	85.4	89.0
KELESC[127]	82.2	78.1	76.7	82.2	83.0	81.2	84.3	69.4	84.0	86.7
Generationary[129]	77.8	73.7	68.8	78.3	77.6	76.3	79.8	63.3	80.1	84.7
Lesk_ext[133]	58.4	59.4	-	-	-	-	-	-	-	-
SREF[134]	72.7	71.5	61.5	76.4	79.5	73.5	78.5	56.6	79.0	76.9
UKB[135]	59.7	57.9	41.7	-	-	-	-	-	-	-
Babelfy[136]	-	68.3	62.7	65.9	-	-	-	-	-	-
SyntagRank[137]	71.6	72.0	59.3	72.2	75.8	71.7	64.1	-	-	-
WSDG[138]	68.7	68.3	58.9	66.4	70.7	67.7	71.1	51.9	75.4	80.9

表 6将上述三类较为特殊的方法放在了由分割线区隔的第一部分，剩下区隔的部分分别对应表 5中的训练过程中的类型。加星号的模型表示采用了更多了语料进行的训练。

2.3.5 问题挑战

词义研究是计算语义学的重要研究课题，计算语言学框架下的词义消歧任务仍旧存在很多问题：

知识论问题着眼于模型的可解释性，类比于“当人类说他掌握一门语言时候，他学习到了什么”，取得较好地消除歧义性能的模型，它学习到了什么知识？模型是否真正掌握或者多大程度地掌握了词汇语义？它与人类可以提炼或者人类直觉的一些语言学知识是否匹配，如果不匹配该如何理解？语言模型是否仅仅学习到某种相关性的模式，而没有学习到一些决定性的、具有因果关系的特征？

与“已知什么”的知识论相对的是，模型是否知道自己不知道什么。在有限条件下（例如残缺的上下文、语域偏移等非正常环境），模型会不会做出过自信的结果，而不会“意识”到自己不知道这方面的知识。现有的最先进的生成式人工智能模型就普遍存在“一本正经地胡说八道”的幻觉（hallucination）问题，这些问题可以使用不确定性进行刻画，本文作者前期工作[169] 将两方面进行结合，做出了一些初步探讨。

任务的定义以及资源的构建。词义消歧任务目的在于全面理解词义的多义性，从而帮助句子语义的理解。如何设计任务去实现它仍然存在争议。主流的分类方法将所有可能的义项当作一个离散的、语义正交的待分类集合，这种定义简单、易操作、却忽略了词义的连续性，以及如何展示“可能的”义项。现有的 WordNet 的词义普遍认为比较精细，这样会导致 ITA 不会处于较高水平。同时主流的分类方法大多是确定式建模，也缺乏对于不确定性、模糊性等因素的考虑。在资源方面，除了知识库的设计外，现在的词义消歧任务也面临“知识获取瓶颈”（Knowledge-acquisition bottleneck）的问题：即训练语料库需要大量的语义标注，这与语义相对于词来讲是隐形的性质相关。这些都需要耗费大量人力物力资源，因此如何高效地词义消歧也是未来研究的方向。

数据分布不均衡与泛化能力。由于表示不同词义的词汇分布非常不均衡：即罕见义项的词汇实例数量要远远小于常见义词汇实例，使得模型倾向于选择最常见意思作为结果。另一方面，受到文本语域偏移的影响，模型对于未见过实例的泛化也存在困难。另一个与泛化能力相关的泛化与组合性相关：模型能否对于未见过的但是通过一些构词规则组成的新词也可以做出判断？这一点尤其与汉语相关，汉语词汇分析性更强，更加具有结构性。

适应于汉语的词义消歧任务。现在主流的消歧任务设计主要是英语，然而中文的语言特点、拥有的资源等都与英文的不同。中文词义消歧任务主要采用 HowNet[22] 作为语义词典，它将语义分解为更小的单位：义素，很多方法[170-171] 利用了知网中的义素及其图结构。也有方法设计中文的语义标注语料库，例如古代汉语相关的[172]，或者以现代汉语词典标注[173] 的等等。中文词内部结构较为明显，且搭配更加紧凑，不少方法的设

计[173-175]也利用了汉语的这些特点。然而，汉语仍然有很多独特的地方没有被给予足够重视，例如汉语的消歧应该作用于语素还是词汇等等，这些都有待于进一步研究。

3 主要研究内容

本节主要介绍文章主要的研究内容，分为形式化定义，可能的研究方案，工作特色以及难点分析。

3.1 形式化定义

假设存在一个词汇空间 $\mathcal{W}$，该空间定义为一个概率空间，存在随机变量 W 代表某个单词 $w_i \in \mathcal{W}$ 出现，其出现的概率为：$P(W = w_i)$。其中，由 M 个随机变量集合定义为句子空间 $\mathcal{S}$ 中的一个随机变量：$S = \{W_1, ..., W_M\}$，其出现的概率记作：$P(S = s_i)$，其中 $s_i = \{w_1, ..., w_M\}$。此外，定义一个意义空间 $\mathcal{Z}$，其空间中的随机变量 Z，为不可直接观测的隐变量，代表词汇 W 的语义。对于 S 中的每个词汇而言，同样存在一个 m 大小的意义变量集合：$S_z = \{z_1, ..., z_m\}$，其中 z_i 代表词汇语义，S_z 代表句子语义。本研究主要关注词汇语义 z_i。

考虑到多义性的普遍存在，假设意义变量 Z 服从多项分布（Multinomial Distribution），即 $Z \sim \mathrm{PN}(p_1, ..., p_N)$，其样本空间表示词汇 w 所对应的 N 个候选语义。这里出于研究的方便，本文默认 Z 为离散型随机变量，即假定给定一个词汇 w 的情形下，可以罗列出它的所有可能的语义候选项 $\mathcal{Z}_w$。进一步地，本文区分它的上下文无关语义集和上下文依赖语义集，上下文无关语义指狭义的词汇语义，即不需通过上下文就可以判断该词汇的可能语义，这些语义通常较为固定、永久而常被列入到词典中；上下文依赖语义非常依赖上下文词汇，一般是临时意义。本文简记 w_i 的上下文为 $c_i = \{w_1, ..., w_{i-1}, w_{i+1}, ..., w_M\}$。其中独立语义情形下，$\mathcal{Z}_w$ 和 c 在已知 w 的条件下独立，通过如下方式建模语义与词汇的关系：

$$P(\mathcal{Z}_w, w, c) = P(\mathcal{Z}_w|w, c)P(w, c) = P(\mathcal{Z}_w|w)P(s). \tag{4}$$

其中，$s = \{w, c\}$。对于上下文依赖语义而言：

$$P(\mathcal{Z}_w, w, c) = P(\mathcal{Z}_w|w, c)P(w, c) = P(\mathcal{Z}_w|w, c)P(s). \tag{5}$$

本文重点关注以下几类不同语义距离的词汇子空间：同形异义词 $\mathcal{W}_\mathcal{H}$、针对于实词的多义项词 $\mathcal{W}_\mathcal{P}$、可逆简单句中的成分 $\mathcal{W}_\mathcal{V}$。它们针对不同的词类或者上下文，更多区别可以参考表格 7。同时，我们定义一个度量函数 $\phi(z_i, z_j)$ 来刻画语义候选项集合 $\mathcal{Z}_w$ 中的元素 i 和 j 之间的语义距离，这一距离反映不同语义选项之间的相关性大小。根据定义，可以存在如下的距离关系：

$$\phi(\mathcal{W}_\mathcal{V}) < \phi(\mathcal{W}_\mathcal{F}) < \phi(\mathcal{W}_\mathcal{P}) < \phi(\mathcal{W}_\mathcal{H}) \tag{6}$$

其中，$\phi(\mathcal{W})$ 表示空间 $\mathcal{W}$ 中两元素间的平均距离。

名称	符号	词类	上下文	举例
同形异义词	$\mathcal{W}_{\mathcal{H}}$	任意	任意	“花”的植物义和花钱义
多义项词	$\mathcal{W}_{\mathcal{P}}$	实词	任意	“头”的身体义和首领义
语义角色	$\mathcal{W}_{\mathcal{V}'}$	任意	可互逆的主谓宾	“十个人”的施事性变化[27]

表 7 不同词汇多义性类别的对比

对于 $S = s$ 中的特定词汇 w_i 而言，其意义变量 Z 的分布常常取决于 w_i 所处的环境集合 $\mathcal{E}$ 中，本文重点关注的环境包括：

- 文本上下文 c_i
- 外部知识 $\mathcal{K}$
- 文化心理因素 $\mathcal{H}$

其中外部知识 $\mathcal{K}$ 又可以细分为语言学知识，包含句法知识、词汇构造等，和外部知识，例如来自百科中的定义，与其他事物的关联等。通过如下的方式来建模外部环境条件下的语义、词汇的联合概率分布：

$$P(z, w|\mathcal{E}) = P(z|w, \mathcal{E})P(w|\mathcal{E}). \tag{7}$$

词义消歧旨在确定目标词汇在上下文中的意义，其目标词通常为名词、动词、形容词、副词四大实词类，环境 $\mathcal{E}$ 通常仅仅涉及上下文 c_i。词义消歧任务可以分为两大类别：判别式和生成式。一类判别式任务假定存在一部语义词典 $\mathcal{D}$，它包含由词汇集 $\mathcal{W}$ 到语义集 $\mathcal{Z}$ 的映射。模型需要从所有候选的词义中确定一个最恰当的词义 $z \in \mathcal{Z}$，其中 $z = \max_i P(z_i|w, c_i)$。另一类判别式任务则不考虑语义词典，直接构造一对包含同一个目标词的不同上下文，来判断二者的语义是否相同，例如 WiC 数据集[176]。生成式任务将 z 视作一个意义语句，即一个词汇序列 $z = \{g_1, ..., g_K\}$，目标是建模并解码出最佳的序列，即，$\max P(g_1, ..., g_K|w, c_i)$。它们的区分举例参考表 8。

3.2 研究方案

本文研究语言模型借助词义消歧任务来评估模型对于词汇多义性的处理能力。首先，我们打算研究语言模型的表征如何、在哪里以及多大程度上反映词汇语义；其次，该表征具体在各个多义性类别上如何表现；在评估准确性的同时，我们也关注词义的不确定性建模。

探测语言模型对于词义的表征。语言模型并没有显示地编码词汇语义，作为计算的最小单元 token，每一层神经网络模型都会输出一个稠密向量来表征该 token，但该表征

[27]前者如：一顿饭吃十个人；后者如：十个人吃一顿饭

表 8 以“头”为例，区分三类不同的词义消歧模型的任务定义

任务类别	输入举例	输出举例
第一类判别式	他是队伍的**头**和 {首领义、部位义等}	首领义
第二类判别式	他是队伍的**头** 和 **头**破血流	意义不同
生成式	他是队伍的**头**	首领

多大程度可以反映该词汇的语义，仍旧是一个问题。尤其在以自回归为目标的大语言模型中，token 中的含义更加不明确。这一部分计划使用常用的 probing（探针）手段来探究不同架构的常用语言模型多大程度反映了相应的词汇语义，由此，也可以得到更好地表示一个词汇语义的表征手段，从而为之后的研究打下基础。

各个多义性类别上的任务设计。本文区分了不同类型的多义性，并针对性地利用或设计相关的任务。目前对于实词的消除歧义的规范任务有很多，但一般不将二者分开，我们计划采用一定的标准划分这两种情况，并且观察语言模型对不同类别的消歧行为有所差异。针对可逆句中的及物动词，我们采用第二类判别方式，即收集汉语中常见的可逆句，并找出不可逆句作为参照。前者的动词标注为 0(代表语义差别最大)，后者标注为 1（代表语义差别最小）。之后再观察语言模型对于它们的检测情况。同时，我们计划从语义角色以及跨语言的角度考虑主语和宾语的施事性和受事性的情况，作为影响动词以及句式的可逆性的重要因素。针对虚词的用法，我们采用案例分析的方式，通过收集“重复义”副词相关的跨语言语料，并利用表征来表示它们的不同特征下的语义距离，绘制为语义地图，并与语言学家绘制的语义地图做比较，从而分析语言模型对于功能词的处理。

词义的不确定性建模。词义的不确定性是除了准确性外的一个重要指标。词汇语义的确定受到很多因素影响，使得模型最终无法确信某一个义项，而在某一些义项中都有选择可能。这一现象在语义众多且往往具有关联的功能词义项中尤为突出。因此每一个义项的选择都应该是一个概率选择的问题。针对现有确定性建模模型的概率倾向“过度自信”的问题，本文设计不同的不同的方式表征词义的不确定性，并分析影响词汇在上下文中含义的几项因素。

3.3 工作特色、难点以及创新点

本工作具有以下特色：

跨学科融合，将人工智能中的计算机领域与传统语言学相结合。本研究通过结合计算机科学中的先进算法与传统语言学的理论基础，推动了跨学科的发展。通过这种融合，我们不仅能够利用人工智能的强大计算能力来处理语言数据，还能借助语言学的深刻理解来评估并提升模型的表现。这种跨学科的方法，不仅丰富了语言学的研究手段，

也为人工智能提供了新的应用场景和研究方向。

设计更加复杂、适用于多语言尤其是汉语的任务。在本研究中，我们特别针对多语言环境进行了复杂模型的设计，特别是对汉语的处理进行了深入的优化。汉语由于其独特的语法和词汇结构，常常对自然语言处理提出特殊的挑战。我们的模型通过对汉语特性的深入研究和优化设计，能够更有效地处理汉语文本，同时也兼容其他语言，使其在多语言环境中具有更广泛的应用价值。

考虑多样的、更适合语言事实的评估指标。我们在评估模型性能时，采用了多样化的指标体系，以更全面地反映模型对语言事实的捕捉能力。这些评估指标不仅涵盖了传统的准确性、召回率等基本指标，还包括一些更适合语言特性的专门指标，如语义不确定性等。这种多维度的评估体系，有助于我们更全面地了解模型的实际表现和改进空间。

同时，工作也具有很多挑战和难点：

首先，深度学习的机理本身就缺乏可解释性。深度学习模型尽管在许多任务中表现出色，但其内部机制往往难以解释。模型的决策过程像一个“黑箱”，使得我们难以理解它是如何得出结论的。这种缺乏可解释性的问题，不仅限制了模型在某些领域的应用，还增加了调试和改进模型的难度。如何提升深度学习模型的透明度和可解释性，是我们面临的一大挑战。

其次，如何使结论更加具有跨语言、跨模型等的普适性。我们在研究过程中发现，不同语言和模型之间存在许多差异，使得一些结论在跨语言或跨模型的情境下可能并不适用。如何确保我们的研究结论具有广泛的适用性，能够在不同语言和模型中通用，是一个重要的研究难点。这需要我们设计出更通用的算法和评估方法，并进行大量的跨语言、跨模型实验验证。

最后，如何将结论带到下游任务中，提高人工智能模型的性能。本研究的最终目的是提升人工智能模型在实际应用中的表现。如何将我们的研究成果有效地应用到实际的下游任务中，如机器翻译、文本生成等，以显著提升这些任务的性能，是一个关键挑战。这不仅需要我们在理论上取得突破，还需要在实践中不断优化和验证模型，使其能够在真实世界中发挥最大效用。

本工作具有如下可能的创新点：

1. 分析出语言模型对于不同程度的词汇多义性现象的识别程度，增强对于语言模型的理解。

2. 利用语言模型对语言学的多义性问题进行自动化研究，预测新的语言现象。

3. 建模语言数据中存在的主观性、不确定性等指标，从而更加反映数据的实际分布情况。

4. 从多语言的角度分析多义性这一普遍的语言现象。

4 词义消歧中的不确定性估计

4.1 摘要

本研究聚焦于词义消歧（Word Sense Disambiguation，WSD）任务，该任务旨在根据特定上下文为多义词确定最合适的词义，对于自然语言理解（Natural Language Understanding，NLU）任务具有重要意义。尽管现有的基于监督的学习模型将 WSD 视为分类问题，并在多项基准测试中取得了突破性进展，但这些模型往往忽略了现实世界数据的不确定性估计（Uncertainty Estimation，UE）。鉴于现实世界数据的噪声性（例如上下文残缺）和分布外（Out-of-Distribution，OOD）（例如语域偏移）特性，对不确定性的量化评估显得尤为重要。本文通过深入研究 WSD 基准测试中的 UE 问题，首先对四种不同的不确定性评分方法进行了比较分析，并验证了模型最终输出层的传统预测概率在量化不确定性方面的不足。随后，本文通过精心设计的测试场景，利用选定的 UE 评分方法对模型在捕获数据不确定性（Data Uncertainty）和模型不确定性（Model Uncertainty）方面的能力进行了评估，结果表明模型虽然能够较为准确地反映数据不确定性，但在模型不确定性的估计上存在低估现象。此外，本文还探讨了影响数据不确定性的多种词汇因素，并从词类、形态学特征、词义粒度和语义关系四个维度进行了详尽的分析。相关代码已在 GitHub 平台[28]公开，以供进一步研究和参考。[29]

4.2 引言

在自然语言理解（Natural Language Understanding，NLU）领域中，对给定上下文中的词语进行词义消歧（Word Sense Disambiguation，WSD）是一项基础性任务。此任务针对的是多义项词（polysemy）或同源异义词（homonymy），旨在根据其周围上下文确定最合适的词义。例如，在句子“Book the hotel, please”和“Read the book, please”中，多义词“book”在第一句话中指“预定义”，而在第二句话中指“书籍义”。词义消歧问题普遍存在于所有语言之中，并且自人工智能（Artificial Intelligence，AI）研究的初期就受到了广泛关注[53]。

现有的监督学习方法将 WSD 视为分类问题[116,118-119,121,177]，通过训练基于神经网络的分类器，在 WordNet 等电子词典资源提供的义项中进行选择。尽管这些方法在 WSD 基准测试上取得了最先进的性能，部分[118] 甚至在准确度上超越了人类标注者之间的一致性估计上限（80%），但它们并未捕捉或衡量不确定性。不确定性估计（Uncertainty Estimation，UE）旨在回答模型对其选择正确选项的确信程度。在现实世界的应用场景中，数据往往带有噪声或偏离训练时候的预期分布，此时模型的不确定性估计对未来的

[28] https://github.com/RyanLiut/WSD-UE

[29] 本工作已经被计算语言学顶会 (Findings of) ACL 2023 录用[169]。

决策尤为重要。例如，之后可以将高不确定性的输入交由人类进行分类。

有趣的是，词义消歧中的“ambiguous”一词本身就具有歧义性：根据《韦氏词典》的解释，它既可以指“由于模糊或不明确而产生的怀疑或者不确定性”，也可以指“能够以两种或更多可能的意义或方式被理解”[30]。传统的处理方法只考虑了它的第二层含义，却忽视了与不确定性相关的第一个含义。实际上，词义选择的过程中存在许多不确定性的情境。本文首先探讨了其中的典型两种：模型不确定性和数据不确定性。模型不确定性源于不同训练模型的结构和参数的变化，这往往可以追溯至样本空间的分布偏移，以至于无法通过增加训练数据的量而完善实验结果。例如数据的语域从小说变到了新闻，或者在训练数据中存在高频词义的数据占绝对多数的长尾分布[112]；而数据不确定性则与数据本身带有缺陷有关，即使数据量充足，也无法获得高置信度的结果，例如词汇的上下文短缺或者不充分。

本文通过大量的实验评估了最先进模型（State-of-the-Art，SOTA）[119] 在 WSD 基准测试中的不确定性。首先，我们比较了模型输出概率与其他三种不确定性评分方法，并得出结论，单一前向传播得到的 Softmax 输出概率不足以进行 UE。接着，我们使用选定的评分方法，在两种设计好的测试场景中评估了数据不确定性。此外，我们还在一个现有的分布外（OOD）数据集[168] 上估计了模型不确定性，并发现模型对模型不确定性的估计低于数据不确定性的适当度量。最后，我们设计了可控实验来确定哪些词汇因素影响不确定性估计。结果表明，形态学（词类、形态单位数量）、词典组织（标注真实词义的数量和多义程度）以及语义关系（下义词）对不确定性评分有影响。

4.3 方法

给定一个目标词 w_i 在上下文 $c_i = (w_0, w_1, ..., w_i, ..., w_W)$ 中的 W 个词，词义消歧（WSD）模型从候选词义集合 $S_i = (y_1, y_2, ..., y_M)$ 中选择最佳标签 $\hat{y}_i$。带有参数 θ 的神经网络 p_θ 通常通过 softmax 函数对模型输出 f_i 进行归一化，得到 M 类上的概率 p_i：

$$p_i = \text{SoftMax}(f_i(w_i|c_i;\theta)) \tag{8}$$

在训练过程中，该概率用于计算交叉熵损失，并在推断时将其视为每个候选义项的概率。这种模型函数的点估计被误解为模型置信度[99]。UE 的目标是找到合适的 p_i，以更好地反映数据和模型不确定性下的真实预测分布。假设我们有一个合理的分数 $s(p_i) \in \mathcal{S}$ 指示 UE，其中 $\mathcal{S}$ 是一个度量空间，当情景 a 比 b 更不确定时，我们期望 $s^a > s^b$。

数据不确定性：可控上下文

数据不确定性衡量由不完美或噪声数据引起的不确定性。我们认为，这些噪声可能出现在目标词周围的上下文中，因为 WSD 是一个上下文敏感的任务。通过控制上下文

[30] https://www.merriam-webster.com/dictionary/ambiguous

中的缺失部分的程度，模型在不同的上下文条件下会获得不同的不确定性预测。为了模拟这种情况，我们基于两个信号控制上下文范围：窗口和句法，如图1所示。

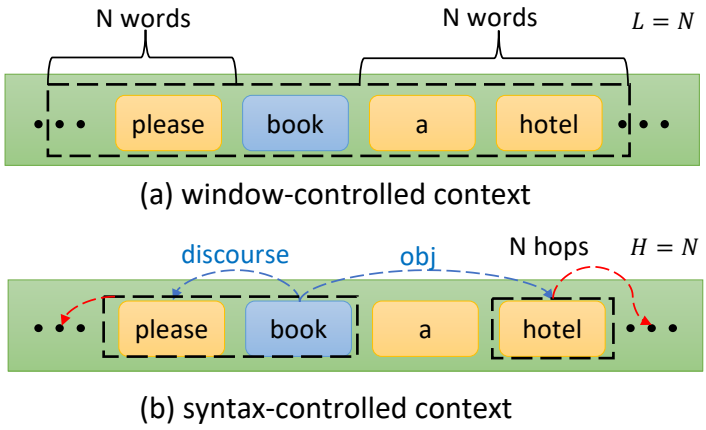

图 1 数据不确定性设置中的两种可控上下文类型。目标词以蓝色高亮显示。黑色虚线框显示最终选择的上下文。蓝色和红色表示句法依存关系。

窗口控制的上下文 我们选择目标词 w_i 左右两侧的 L 个词作为窗口控制的上下文

$$c_L^{\mathrm{WC}} = (w_l, w_{i-1}, w_i, w_{i+1}, ..., w_h),$$

其中 $l = \max(i - L, 0)$ 和 $h = \min(i + L, W)$ 分别是下限索引和上限索引。假设较长的上下文往往包含更多的词义消歧线索，并且合适的 UE 分数 s，我们期望 $s_a^{\mathrm{WC}} > s_b^{\mathrm{WC}}$，其中两个窗口控制的上下文长度分别为 a 和 b，且 $a < b$。

句法控制的上下文 在我们的第二种控制方法中，我们利用 w_i 周围的邻近句法。具体来说，我们使用 Stanza 工具解析词之间的通用句法依存关系。其表示为图结构 $\mathcal{G} = (\mathcal{N}, \mathcal{R})$，其中 $\mathcal{N}$ 表示节点，即每个词，$\mathcal{R} =< n^h, n^t, r >$ 表示从头节点 n^h 到尾节点 n^t 的关系 r。例如，当 r 是 *nsubj* 时，表示 n^h 是 n^t 的主语。我们通过以下方法迭代获得包含 H 跳的目标词 w_i 的句法相关邻近集 c_H^{DP}。最初，c_H^{DP} 只包含 w_i。经过一跳后，c_H^{DP} 收集 w_i 的头节点和尾节点。该过程重复 H 次，添加更多句法相关的词。我们合理假设测量句法控制上下文下的不确定性的 s^{DP} 较小，有利于包含较大 H 的上下文。需要强调的是，句法控制的上下文利用词之间的非线性依存距离[178]，而窗口控制上下文中的距离是线性的。

模型不确定性：OOD 测试

模型不确定性是 UE 的另一个重要方面，广泛研究于机器学习社区。由于知识的缺乏，不同架构和参数的模型可能输出不确定的结果。在 OOD 数据集上测试模型是评估模型不确定性的一种常用方法。在 WSD 任务中，我们使用现有的数据集 42D[168] 作为更具挑战性的基准。这个基于英国国家语料库的数据集具有挑战性，因为 1）每个实例的真实值不在 SemCor[106] 中，后者是 WSD 的标准训练数据，2）避免了最频繁的词义

偏差问题，42D 的词义不在 WordNet 的第一个义项中[179]。此外，42D 的数据域与训练语料库不同。这些特点使得 42D 成为理想的 OOD 数据集。

4.4 实验

模型和数据集 采用在词义消歧任务上目前性能优越的模型 MLS[119] 进行实验，这个模型将 WSD 重新定义为一个多标签问题，并在标准 WSD 训练数据集 SemCor[106] 上训练了一个 BERT-large-cased 模型[1]。本次实验不改变它训练时候的参数设置，但在推理过程中更改随机失活机制 Dropout[99]，从而来执行蒙特卡罗 Dropout（MC Dropout）。我们将样本数量 T 设置为 20，进行 3 轮实验，并报告平均性能。评估基准使用一个统一的评估框架[112]，它包括五个标准数据集，即 Senseval-2、Senseval-3、SemEval-2007、SemEval-2013 和 SemEval-2015。在第二部分本文仅使用 SemEval-2007 来研究数据不确定性，并使用 42D 来研究模型不确定性。

不确定性估计分数 采用四种方法作为不确定性估计（UE）分数。一种简单的基线方法[180] 将 Softmax 输出 p_i 视为类 $y = s \in S$ 的置信度。基于此，不确定性可以表示为 $u_{\text{MP}}(x) = 1 - \max_{s \in S} p(y = s|x)$。

另外三种方法基于 MC Dropout，即在推理过程中通过 Dropout 随机掩码进行 T 次随机前向传递，得到 T 个分类概率 p_t[181]：

- 采样最大概率（SMP, Sampled maximum probability）将样本均值作为最终置信度，然后应用 MP：$u_{\text{SMP}} = 1 - \max_{s \in S} \frac{1}{T} \sum_{t=1}^{T} p_t^s$，其中 p_t^s 表示第 t 次前向传递时属于类 s 的概率。
- 概率方差（PV, Probability variance）[182] 在对所有类概率求均值之前计算方差：$u_{\text{PV}} = \frac{1}{S} \sum_{s=1}^{S} \left(\frac{1}{T} \sum_{t=1}^{T} \left(p_t^s - \overline{p^s} \right)^2 \right)$。
- 通过分歧进行贝叶斯主动学习（BALD, Bayesian active learning by disagreement）[183] 测量模型参数与预测分布之间的互信息：$u_{\text{BALD}} = -\sum_{s=1}^{S} \overline{p^s} \log \overline{p^s} + \frac{1}{T} \sum_{s,t} p_t^s \log p_t^s$。

UE 分数的评价指标 虽然 UE 分数是一种不确定性度量，但需要指标来判断和比较不同 UE 分数的质量。假设具有高不确定性分数的样本更可能是错误的，移除这些实例可以提升性能。我们采用文献[181] 的两种度量方法：风险-覆盖曲线面积（**RCC**）[184] 和不一致对比例（**RPP**）[185]。RCC 根据预测拒绝的不确定性水平计算分类错误的累计损失。较大的 RCC 表明不确定性估计对分类产生负面影响。本文使用通过数据集大小归一化后的 RCC。RPP 计算不确定性水平与其损失水平不一致的实例比例。对于任意实例对 x_i 和 x_j，其 UE 分数 $u(x)$ 和损失值 $l(x)$ 满足以下公式：

$$RPP = \frac{1}{n^2} \sum_{i,j=1}^{n} \mathbb{1}[u(x_i) < u(x_j), l(x_i) > l(x_j)], \tag{9}$$

其中 n 是数据集的样本大小。

4.5 结果

本小节主要从定量角度分析哪个 UE 评分更加适合；之后从定性角度展示不同不确定性程度的多义词，最后分析影响不确定性评估的因素。

4.5.1 定量分析

整体上分析哪个 UE 评分更好。我们使用 RCC 和 RPP 两个指标来衡量 MP、SMP、PV 和 BALD 四种不确定性评分的表现。表9展示了五个标准数据集上的结果，而涉及不同词类的数据集的性能则在表10中展示。对于大多数数据而言，尽管 MP 在某些情况下（如在 SemEval-15 上）稍有优势，SMP 表现更加出色，超过了其他三种评分。有趣的是，基于 SoftMax 的 MP 和 SMP 评分优于 PV 和 BALD。类似的结果在文献[181] 中也有所观察到。这可能是因为前者的评分直接用作最大似然目标的输入，因此更准确地逼近真实分布。

表 9 五个标准 WSD 数据集上的 UE 分数比较。

UE 分数	Senseval-2		Senseval-3		SemEval-07		SemEval-13		SemEval-15	
	RCC ↓	RPP ↓	RCC ↓	RPP ↓	RCC ↓	RPP ↓	RCC ↓	RPP ↓	RCC ↓	RPP ↓
MP	**5.69**	9.50	7.11	10.37	**8.68**	11.40	5.78	8.02	**5.02**	**11.07**
SMP	5.78	**9.14**	**7.10**	**9.83**	8.81	**10.83**	**5.59**	**7.88**	5.34	11.16
PV	6.11	11.47	7.50	12.40	9.93	16.00	5.97	10.22	5.62	13.11
BALD	6.00	11.09	7.46	11.99	9.36	14.73	5.83	10.02	5.48	12.77

为进一步探究这四种评分的分布情况，我们展示了它们在误分类实例中的直方图，如图2所示。我们还展示了平均值（红色虚线）和样本偏度 s 的计算结果。结果显示，MP 的分布更长尾和偏斜，比基于 MC Dropout 的评分表现更为过于自信，这验证了单次前向 SoftMax 输出作为置信度指标的普遍问题。

表 10 不同词性的数据集上 UE 分数的比较。

UE 分数	名词		动词		形容词		副词		全部	
	RCC ↓	RPP ↓	RCC ↓	RPP ↓	RCC ↓	RPP ↓	RCC ↓	RPP ↓	RCC ↓	RPP ↓
MP	6.06	**7.47**	14.08	18.20	5.15	**8.25**	3.70	4.89	6.13	9.78
SMP	**4.94**	7.66	**13.76**	**17.45**	**4.39**	8.35	**2.65**	**4.85**	**6.11**	**9.44**
PV	6.25	9.17	15.38	22.02	4.97	9.37	3.20	5.33	6.48	11.91
BALD	5.18	9.39	14.42	20.96	4.59	9.80	2.66	5.56	6.36	11.52

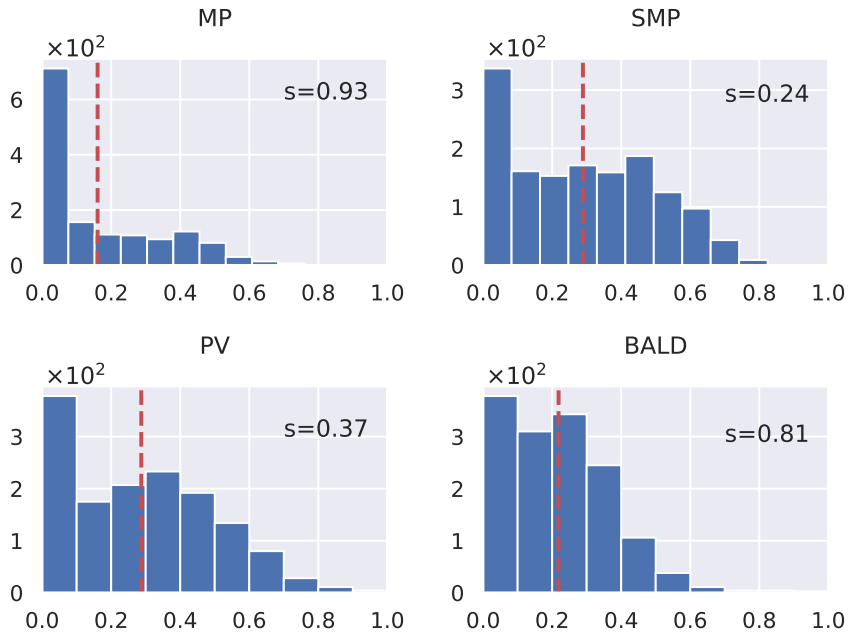

图 2 数据集中误分类实例上四种 UE 评分的分布情况

最后，鉴于其出色的表现，后续实验选择 SMP 作为接下来的不确定性评分。

模型反映数据不确定性的情况。在窗口控制和语法控制场景中验证了数据不确定性，如图3所示。在第一个设定中，随着窗口大小 T 的增加，不确定性分数下降且准确性提高。这表明模型在接触到更多相邻词汇时，对数据的确信度逐渐增强。在语法控制设定下，趋势类似。这些结果表明模型能够充分捕捉数据不确定性。在稀疏上下文中（例如 L 或 H 等于 0 或 1），SMP 的不确定性比分数明显大于 MP，模型在这些情况下的不确定性更大。

模型反映模型不确定性的情况。图4中考察了 42D 数据集上的模型不确定性。结果显示，OOD 数据集确实是 WSD 的一个具有挑战性的基准。然而，即使性能较差，模型也未能给出高的不确定性分数。我们将其与数据不确定性设定下最不确定但准确性相近的情况进行了比较，即 $L = 0$ 时没有任何上下文的情况下。即使在性能下降的情况下，OOD 设定的不确定性水平较低，尤其是在错误分类的样本中。这表明模型低估了模型不确定性中的不确定性水平。

4.5.2 定性分析

为了研究在给定上下文的情况下，哪些词语倾向于表现出不确定性，我们通过对共享相同词形的实例平均 SMP 分数来获得每个词的最终不确定性分数。在图5中，我们展示了最不确定（左（a））和最确定（右（b））意义的词语词云。我们移除了一些候选词义少于 3 的无代表性词语。对于最不确定的词汇，有诸如 *settle*、*cover* 等词语，大多数是动词，并且具有多个候选词义。而在最确定的情况下，像 *bird*、*bed* 和 *article* 这样的名

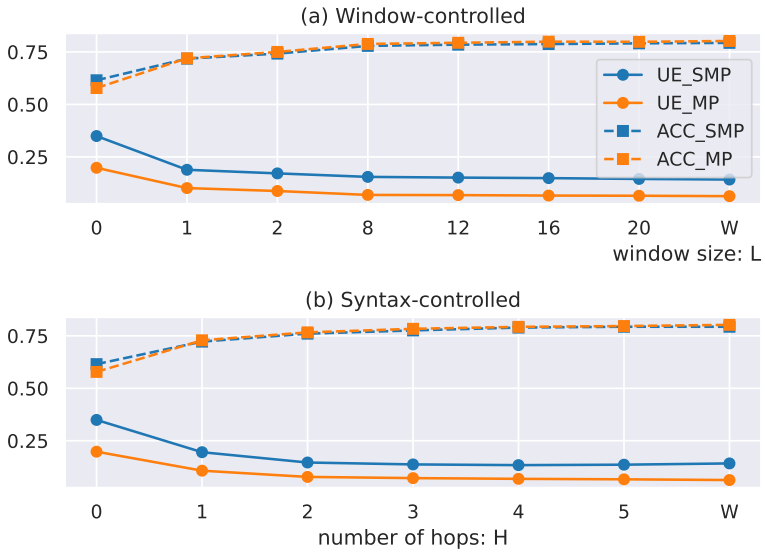

图 3 在 (a) 窗口控制设定和 (b) 语法控制设定下，不同上下文范围对不确定性分数（SMP 和 MP）及准确性（F1 分数）的影响。“0”表示模型只使用目标词而无上下文，“W”表示使用完整上下文。

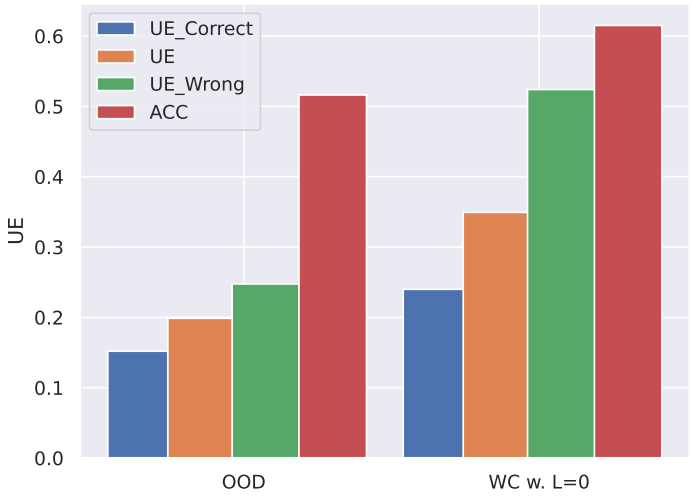

图 4 模型不确定性（OOD）和数据不确定性（受控上下文）场景下的不确定性和准确性（F1）分数。我们使用窗口控制的不确定性分数，设定 L=0（WC w. L=0）。结果在所有数据实例以及错误分类（UE_Wrong）或正确分类（UE_Correct）实例中进行了评估。

词词义不确定性较低。这些现象激励我们在下一部分中研究哪些词汇特性影响（数据）不确定性估计。

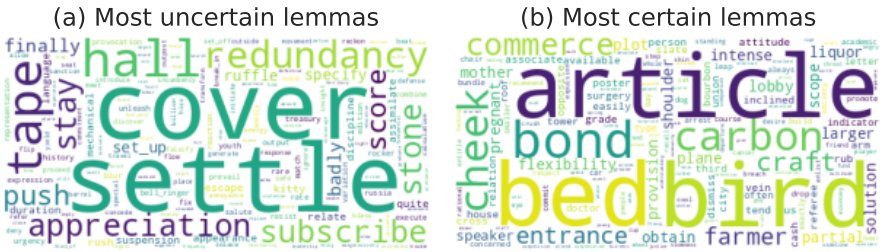

图 5 词形的词云图，其中字体较大表示较高（a）或较低（b）的不确定性分数。

4.5.3 影响因素

考虑到语言学和认知相关的研究，文章从四个方面探讨哪些词汇特性影响不确定性估计：词法类别[186]、形态学[31][187]、词义粒度和语义关系[9]。词法类别关注目标实词的词性（**POS**）。形态学关注词素的数量（**nMorph**）。语义粒度考虑标注的真实词义数量（**nGT** 和候选词义数量，即多义性程度（**nPD**）。语义关系方面关注 WordNet 中的下位词和同义关系。每个词在 WordNet 中作为一个节点位于下位词树中，其深度表示具体化程度，记为 **dHypo**。同时还考虑真实词义所属的同义词集合的大小（**dSyno**）。

为了探究它们的影响，首先对于样本中的各个影响因素按照值的大小进行分层处理，之后我们观察不同水平的影响因素在不确定性分布上是否具有显著差异。分层结果可以参考表11。

热力图 6显示了不同的**词类**对于不确定性的影响。除了名词-形容词组合外，动词实例比名词或形容词更显著地不确定，而副词的不确定性最小。这一结果表明，动词的语义一般比其他类别更难确定，与先前的研究结果一致[126,179]。这一点可以从表 10 和图 5 中得到验证。

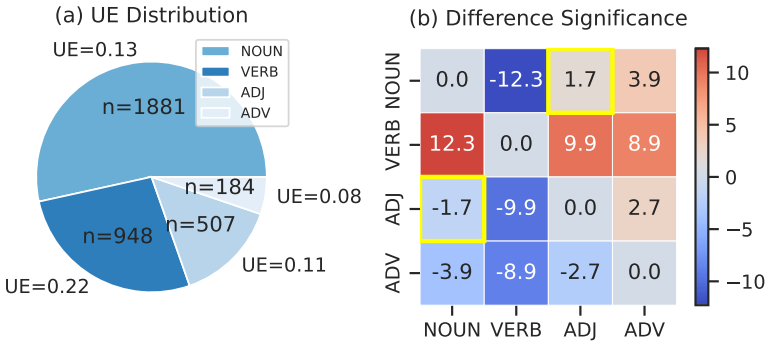

图 6 按词义聚合实例的平均不确定性得分及其数量，分别按不同词性分类 (a) 以及每对词性的差异显著性 (b)。热图 (b) 显示了 T 检验值，其中绝对值越大（颜色越深的格子）表示差异越显著。我们突出显示了对应 p 值大于 5% 的格子，它们没有显著差异。

[31]这里我们主要考虑派生形态学。多词表达（例如复合词）也包括在内，具有不同屈折形态的词被视为相同的词形。

表 11 针对不同影响因素，不同层级的数量和范围。

影响因素		L1	L2	L3
nMorph (N)	数量	514	603	397
	范围	(0,1.67]	(1.67,2]	(2,9]
nMorph (V)	数量	200	313	132
	范围	(0,2)	[2,2]	(2,6]
nMorph (A)	数量	136	201	69
	范围	(0,1.30]	(1.30,2]	(2,6]
nMorph (D)	数量	25	85	36
	范围	(0,2]	[2,2]	(2,6]
nGT	数量	6913	340	-
	范围	1	>1	-
nPD	数量	1145	963	463
	范围	(0,2]	(2,6]	(6,50]
dHypo	数量	729	666	340
	范围	(1,6]	(6,9]	(9,43]
dSyno	数量	1109	1407	763
	范围	(0,1]	(1,3]	(3,28]

表 12 不同层级的不同不确定性估计（SMP）以及涉及形态学、词义库组织和语义关系的各种影响因素的相应差异显著性（p 值）。Agg. 表示词形（L）、实例（I）和词义（S）的聚合方式。

影响因素	条件	聚合方式	不确定性估计			差异显著性		
			L1	L2	L3	L1 ↔ L2	L1 ↔ L3	L2 ↔ L3
nMorph	nGT=1, POS=NOUN	L	0.13	0.11	0.07	**1.44e-2**	**1.35e-8**	**5e-4**
	nGT=1, POS=VERB		0.22	0.19	0.13	7.61e-2	**6.04e-4**	6.6e-2
	nGT=1, POS=ADJ		0.11	0.08	0.10	**3.6e-2**	4.21e-1	4.40e-1
	nGT=1, POS=ADV		0.11	0.06	0.02	7.6e-2	**6.04e-4**	6.60e-2
nGT	-	I	0.12	0.22	-	**1.61e-22**	-	-
nPD	nGT=1	L	0.04	0.16	0.22	**6.22e-96**	**3.42e-135**	**5.01e-10**
dHypo	nGT=1, POS=NOUN	L	0.14	0.12	0.09	**1.43e-2**	**1.91e-6**	**6e-3**
dSyno	nGT=1	S	0.14	0.14	0.14	5.55	5.38	5.67

表 12展示了剩下的几类情况对于不确定性的影响。从**词汇形态学**角度上看，词语包含的词素越多，其语义就越不确定。从派生形态学的角度来看，这是可以预料的，因为添加前缀或后缀可以特定化词干词，并具有相对可预测的含义。例如，“V-ation” 表示

词干动词的动作或过程，例如 education（教育）、memorization（记忆）。根据表 12 中的 T 检验结果，名词不同级别的 UE 分数显著不同，而其他类别的差异则不那么显著。这是因为包括复合词在内的派生名词比其他类别更具代表性和能产性。这可以通过表 11 中显示的名词包含最多词素的事实来证明。

从**词义粒度**的 nGT 角度，在注释过程中，约 5% 的目标词被标记为多义词义[119]。这反映了即使对于人类注释者来说，选择最合适的含义也是困难的。考虑到它们的上下文，这些词的语义选择更不确定，我们的结果与事实一致。我们在剩余的评估中控制 nGT 为 1，以消除其影响。其次，我们研究了多义度 nPD 的影响。结果表明，具有更高多义度的目标词倾向于更不确定。这在直观上是可以理解的，因为具有更多可能含义的词通常更常见，更容易发生语义变化，例如 *go*（去）、*play*（玩）。此外，它们在 WordNet 中的词义描述更为精细，有时即使对人类来说也难以区分。然而，像复合词这样的低多义度词在各种上下文中更为确定。

从语义关系角度上，首先考虑到上义词关系，即一个词节点在上义词关系树中的深度（dHypo）。由于名词具有更清晰的上义词关系实例，我们只考虑这一类别。表12显示的结果表明，具有更深上义词的实例往往拥有确定的含义，且每对级别之间的差异都显著。这表明更具体的概念具有更明确的消歧能力。另一个语义关系是同义词关系，用 dSyno 表示。测量结果显示，不同同义词数量级别之间的实例在语义上并没有显著差异。这意味着，无论真实的含义是否有更多具有相似语义的实例，对不确定性的决定影响较小。

4.6 结论

我们探索了词义消歧（WSD）的不确定性估计。首先，我们比较了各种不确定性评分。然后，我们选择了 SMP 作为不确定性指标，并检查了当前最先进模型在捕捉数据不确定性和模型不确定性方面的能力。实验表明，该模型能够充分估计数据不确定性，但低估了模型不确定性。我们进一步探讨了从形态学、词义库组织和语义关系等角度影响不确定性估计的因素。未来，我们将在下游应用中将词义消歧与不确定性估计整合起来。

5 大语言模型表征对词义的反映

5.1 摘要

大型语言模型在通用语言理解任务中取得了显著的成功。然而，作为一类生成方法，其目标是预测下一个标记，与其前辈（如 BERT 等架构）不同，这些模型的语义演进深度尚未完全探索。本文针对一种流行的大型语言模型 Llama2，通过在每个层的

末端探测其隐藏状态进行了特定的自底向上词汇语义演变研究，采用了上下文中词语语义是否相同的识别任务。我们的实验表明，较低层的表示编码了词汇语义，而具有较弱语义归纳能力的较高层则负责预测。这与具有判别目标的模型（如掩蔽语言建模）不同，后者的高层获得了更好的词汇语义。通过在提示策略中最后的无意义符号（如标点符号）的隐藏状态上的单调性性能增加进一步支持了这一结论。我们的代码可在 https://github.com/RyanLiut/LLM_LexSem 获取。[32]

5.2 引言

近年来，以生成式预训练语言模型（GPT）为基础的大型语言模型（LLMs）在各种理解和生成任务中展现出令人印象深刻的性能，这种方法不同于之前以 BERT 模型为底座的预训练-微调。然而，现有研究指出[189]，GPT 模型的上下文表示在下游任务中表现不佳，难以完全捕捉单词的语义细微差别。这种差异引发了一个关键的研究问题：LLMs 在多大程度上以及通过哪些层去编码词汇语义？

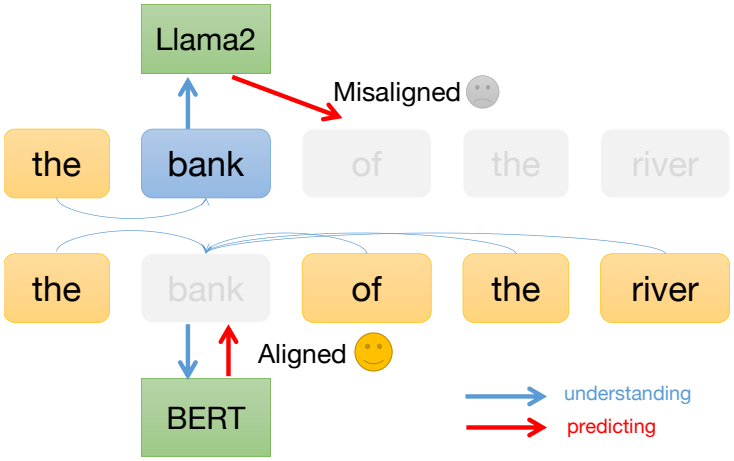

图 7 BERT 和 Llama2 语言模型的关键差异。蓝色和红色线条表示理解和预测的信息流。这里的“理解”指利用上下文捕捉词汇语义。从上下文到当前词的蓝色线条表示理解的流向。

对 BERT 中间层输出的表示进行的研究揭示了模型可以编码一些重要的语言信息，包括其层次结构[190]。例如，BERT 在底层编码表面特征，在中间层编码句法特征，在顶层编码语义特征。然而，由于结构上的差异和挑战，LLMs 中的上下文表示受到了较少的关注，理由可以参考图 10。首先，LLMs 采用了只有解码器的策略，在推断期间仅限于访问目标词之前的上下文（即上文）。因此，LLMs 在相同上文时候无法区分词义。例如“bank”的“河岸”和“银行”共享了左侧上下文“the”。此外，LLMs 训练的目标是预测下一个符号（token），导致在不同层上对历史的和预测的上下文的理解程度不同。相比之下，BERT 专注于通过掩码语言建模（MLM）进行掩码词恢复，同时针对同一个

[32]本工作已经被计算语言学顶会 (Findings of) ACL 2024 录用[188]。

词进行理解和预测，也就是这两个过程是一致的。

基于这些观察，我们假设 LLMs 在较低层编码词汇语义，在较高层遗忘与当前目标词相关的信息，而转而关注下一词的信息。[33]这种层次化行为表明，在生成式 LLMs 中理解和预测信息之间存在动态交互，如最近研究中信息流的观点所示[191-192]。

为验证我们的假设，本研究深入分析了 LLMs 中的词汇语义，通过分析每一层隐藏状态如何反映词义。具体而言，我们利用词语在上下文基准测试中的理解，考察了流行的开源 LLM——Llama2[193] 中的词汇语义。我们采用了各种输入转换和提示语策略来充分利用上下文信息。结果表明，Llama2 的较低层捕捉了词汇语义，而较高层更加重视预测任务。这些发现为决定在生成式 LLMs 中使用哪些层次的隐藏状态作为当前词义的表示提供了证据。

5.3 实验设计

5.3.1 探测任务

我们利用 Word in Context（WiC）数据集作为探索词汇语义的探测任务[176][34]。它是一个二元分类的任务：判断相同的词在不同语境中是否传达相同的含义。其中 638 个实例构成的开发集用于微调最佳的参数，并在测试集的 1400 个实例上评估最终性能。最终计算整体以及不同词性上的准确率。

5.3.2 实验设定和模型

对于出现在上下文 c 中的目标单词 w[35]，由 32 层 Transformer 构成的大语言模型 Llama2 针对每一层 i 来提取表征 $h_i \in \mathbb{R}^D$，其中表征的维度 D 为 4096。之后计算单词 w 在不同上下文（c^a，c^b）中的余弦相似度 s_w^{ab}。随后，如果 s_w^{ab} 超过阈值 γ，即认为单词 w 在两类上下文中的语义一致，分类结果为真，如果低于 γ，则分类为假。我们通过开发数据集确定最优 γ，它对于每一层都应该是不同的，这些值最终列在附录 1.1 中。之前有研究表明[189]，表征空间上的点分布在一片集中区域，而非均匀散布在空间中，为了避免这一现象，我们采样了样本标准化方法。

我们设置不同的输入格式来评估 Llama2。**base** 设置使用原始上下文 c 和目标位置的词汇表示 h_i。在这种设置中，由于在自回归模型中 w 无法访问其后的上下文，我们将原始上下文重复一次，并在第二个上下文中获取 w 的表征。这种配置称为 **repeat**。此外，还探索了目标单词前面的单词，以验证更高层的预测能力，称为 **repeat_prev**。另一种设置灵感来自于文献[194] 中提出的提示语策略，即修改上下文 c 为：*The w in this sentence: c means in one word :*。然后，我们从最后一个标记的位置即冒号 **:** 计算表征 h_i，并将其

[33]较低层指靠近输入层，较高层指接近输出层。

[34]https://pilehvar.github.io/wic/

[35]我们将单词内所有 token 的表征做平均后再作为最终单词的表示。

表 13 输入设定以及输入实例，加粗位置表示目标词的表征提取的位置。

设定	输入示例
base	the **bank** of the river
repeat	the bank of the river the **bank** of the river
repeat_prev	the bank of the river **the** bank of the river
prompt	In this sentence “the bank of the river”, “bank” means in one word **:**

表 14 WiC 测试集上整体的准确率。† 表示的是没有使用样本归一化的结果，括号里面的数字表示模型达到最优结果时候的层数。

方法	所有实例	名词实例	动词实例
人类基准	80.0	-	-
随机选择	50.0	-	-
WSD	67.7	-	-
BERT_large†(23)	67.8	69.1	67.6
BERT_large (22)	71.0	70.7	71.5
Context2vec	59.3	-	-
Elmo	57.7	-	-
Llama2_base†(6)	60.9	63.7	58.3
Llama2_base (11)	63.6	66.8	58.7
Llama2_repeat†(9)	64.5	66.4	63.4
Llama2_repeat (8)	68.1	72.7	65.6
Llama2_prompt†(28)	71.1	68.9	**72.9**
Llama2_prompt (21)	**72.7**	**74.5**	72.1

命名为 **prompt**。表 13 展示了一个例子 [36]。

为了比较自回归生成模型与双向模型，我们在 BERT-large[37] 上进行实验。该模型为 25 层，隐层维度为 1024 和参数量为 336M。此外，我们还考虑了其他词级别的上下文表征的方法，例如 WSD[195]、Context2vec[196] 和 Elmo[197]，结果来自原始数据集论文[176][38]。

5.4 结果分析

表 14 展示了**总体性能**。Llama2 作为生成模型的结果优于双向和非回归模型 BERT 以及静态模型 Elmo，这表明虽然 LLMs 并没有显示学习词义的信息，它在词级理解方面具有巨大潜力。其中，提示语策略在所有 Llama2 的所有设定中的准确率最高。这种方法在将词义理解转化为 LLMs 训练时候的生成模式，已经被证明更加有效[198]。然而，提示依赖于提示词的选择，可能无法直接揭示模型的内部理解。相对而言，重复策略表现得与提示语策略相当，并显著优于基础版本（优势差距为 4.5）。这种简单有效的转换在信息可达性和提示鲁棒性之间取得了平衡。

在词性方面，名词通常比动词表现出更高的准确率，如在 Llama2_repeat 中名词的优势差距为 7.1。我们还观察到，在基础设置中，动词的准确率显著较低，比名词减少了 8.1 个百分点。这是因为动词的消歧需要更多的上下文，而在动词前缺乏上下文的真实数据中，这种上下文往往不足。例如，19.2% 的动词示例位于句子的开头，它们由于没有获得有效的上下文而无法消除歧义，而这一情况名词占比为 14.3%。这些观察结果与之前研究得出的动词更难消歧的结论一致[126]。

之后我们探究了一些消融实验。**首先**是归一化确实起到了非常显著的效果，在每一种设定上性能都得到了有效的提升。**其次**是跨层之间的趋势。图 8 展示了 Llama2 和 BERT_large 在两种设置下的不同层之间性能的动态变化。其中 Llama2 在各层中表现出非单调趋势：base 和 repeat 设置在较低层中表现逐渐上升，然后在较高层中下降。因此，当使用目标词的隐状态作为默认选择时，在较低层可以实现最佳性能。这表明 LLMs 的较低层编码了词汇语义。这一趋势与双向的 BERT_large 模型形成对比，后者在较高层中获得最佳性能。这突显了这两种架构之间的差异：BERT 在各层中集中于当前词，而 Llama2 则着眼于下一个词的预测。

最后是关于理解和预测的平衡。我们观察在 **repeat_prev** 的表现，它是目标词前一个词的表征。需要注意的是，我们选择了 repeat 设置而不是 base 设置，因为 base 设置受限于不完全的信息访问。此外，我们还与 prompt 设置进行了比较，如图 9 所示。尽管这些表示并非来自正确的目标词，而是它的上一个词，但 repeat_prev 和 prompt 在各层中表现出单调趋势和相近的结果。这一观察表明，尽管随着层数加深理解能力可能会减弱（如 repeat 设置中的倒 U 形趋势所示），但预测能力有所提升。

[36]为了方便起见，这里显示为英语例子

[37]https://huggingface.co/bert-large-uncased

[38]值得注意的是，本文复现的 BERT-large 结果略高于数据集论文中报告的性能。

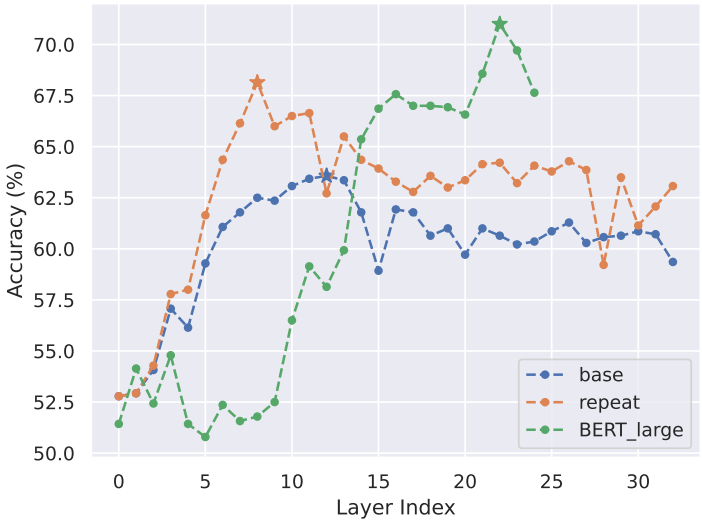

图 8 Llama2 模型的不同输入设定下的性能随层数的变化。星表示最佳性能。

5.5 结论

本研究利用 WiC 数据集调查了 Llama2 为代表的大语言模型中各层表征如何编码词汇语义。实验结果表明，较低层编码词汇语义，而较高层进行预测。这表明 Llama2 在信息从低层流向高层时先进行理解，再进行预测。这些发现为工程应用中与词汇语义相关的任务提供了实用指导。例如，对于词性标注和词义消歧等词汇相关任务，我们应选择使用较低层的表示；而对于文本摘要和对话生成等预测或生成任务，则可以使用较高层的表示。此外，这些发现还从自上而下的角度为大语言模型的可解释性提供了新的视角。

6 汉语主谓宾句主宾互易数据集构建和评估

6.1 摘要

作为典型的形态不发达语言，汉语的基本语序，即主谓宾结构，编码了句子语义的重要方面，其典型的语义表示为施事者（主语）对受事者（宾语）施加了某种行为（谓语）。在一般情况下，主语和宾语易置，会导致语义相反（例如“张三打李四”和“李四打张三”）甚至语义不合规范（例如：“他踢石头”和“石头踢他”）。然而汉语仍存在不少主宾互易后合法且基本语义未发生变化的语言实例，例如“行人走便道”和“便道走行人”。这样就构成了形式上相同（主宾互易），但语义变化的三类最小对立对：语义不变，语义相反，语义不合规范。由于语序在汉语的表义发挥的重要作用，捕捉这几类情况的语义变化对于理解现实世界至关重要。大语言模型在海量数据中进行训练，并取得了卓

图 9 Llama2 模型的不同输入设定下的性能随层数的变化。

越的文本理解能力。一个值得探究的问题是，它们是否能够隐式地学习到语序变化对于不同句子对的语义变化的差异？同时模型是否可以感知到不同成分的语义角色变化？为了对这个问题进行研究，本文构建了一个大规模的汉语简单句的主宾互易数据集，它由上述三类的最小对立对构成，即语义不变，语义颠倒，语义不合规范。之后我们将该数据集定义为一个文本语义相似度任务，对大语言模型在内的多个模型进行详尽的评估。[39]

6.2 引言

语序特征对于词汇形态不发达的汉语的语义表达中发挥着重要作用。汉语的基本语序编码为“主谓宾结构”（Subject-Verb-Object, SVO），其典型的语义可以表达为：主体（施事者）对客体（受事者）实施了某种行为，例如：“张三打李四”。在这类情况下，主宾的顺序互换会因主客体的交换而语义相反[40]，从而给听者带来巨大的语义差异（“李四打张三”），从而影响他的信息接收和对客观世界的理解。在某些情况下，由于动词的语义搭配限制，互易后的句子甚至不合当前语义规范（anomaly），成为无所指陈的句子[41]。例如：“他踢石头”和“* 石头踢他[42]”在这类情况下，汉语的语序清晰地编码了施受关系：处于前面的论元（S）编码了施事者，处于后面的论元（O）编码了受事者。

然而汉语的语序并非严格对应这种施受角色，例如大量的受事主语句就是一个有力

[39] 本部分研究还在进行中，仅显示了已经完成的部分。

[40] 本文所知的语义相反是指主客体颠倒而造成听者对于信息理解上的相反。

[41] 该类句子中的一个著名的例子来自乔姆斯基：Colorless green ideas sleep furiously[199].

[42] 本文用 * 表示语义不合规范的句子。

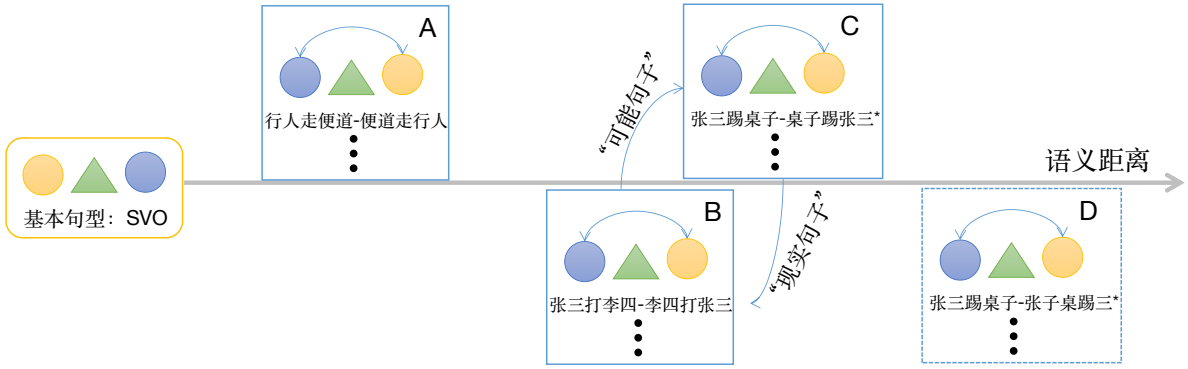

图 10 主宾语交换的四类句对以及交换后的句子距离原来句子的语义距离差距。其中 A 类表示主宾交换语义基本保持不变，B 类表示主宾交换语义相反，C 类表示主宾交换后语义不合规范，D 类是一个对照组，是句内字之间相互交换。其中 C 类句子受限于情境和现有的语义规范成为相对于 B 类现实句子的“可能句子”。图中的黄色圆圈、蓝色圆圈和三角形分别表示主语、谓语和宾语。

的证明[200]。就 SVO 的基本句式而言，就存在不少主宾语相互交换而语义保持基本不变的现象[201]。例如，“十个人吃一顿饭 $\Longrightarrow$ 一顿饭吃十个人”。这时，变易后的相同成分的语义角色往往不发生改变：“十个人”在前后变化前都是施事者，“一顿饭”在前后都是受事者，后一句构成了一个受事主语句。上述句子尽管主宾顺序交换，却不会给听话人造成理解上的干扰，这一现象受到语言学的广泛关注和研究[36,201-204]。

上述三类互易的句子构成了很明显的对照，我们在图 10中呈现了它们的区别。在每一组句子对中，由于仅仅有主宾成分的顺序发生了变化，其他保持不变，因此它们构成了一个最小对立对（minimal pair）。交换后的句子距离原来句子之间的语义距离也呈现出相应的差异。通过最小对立对，我们可以清晰地感知到语序对于不同类型句子的不同影响。

当前的大语言模型已经在各类自然语言理解任务上取得卓越体现，但仍需要对它的语义理解能力做更加精细的评估。评估模型是否可以捕捉上述三类互易句语义变化情况非常重要：一方面，通过大规模语料库进行学习的语言模型基于语义的分布式假设[205]，只关注语言材料本身的表层方面（surface level），而上述类型体现了相同的形式下（即都是主宾变化），不同的语义变化，正确理解不同的变化需要模型从深层的结构和语义的角度（deep level）对句子进行理解。另一方面，对 SVO 各个成分的语义关系的捕捉是理解更多复杂的、自然产生的句子的基础，因此它构成了认知范畴中原型施事和原型受事一对基本概念，并且广泛存在于各种语言之中[13]。汉语的语序变化作为这类普遍概念中的参数（parameter），体现了特殊的结构和语义特点。检验模型这方面的评估能力，对于理解这一广泛存在的认知范畴也有重要的启示。

为了针对这一问题进行研究，我们构建了中文主宾互易对的评测数据集。该数据集包含三组句对，第一组是主宾语交换成分，基本语义不发生变化的；第二组是交换后语义发生了主客体语义相反；第三组则交换后语义不符合规范。数据集的收集采用半自动

的方式，首先通过启发式的方式设计一些典型的基础句子，之后再根据外部语义词典对各个成分进行语义增强。最后，人工进行校验审核。最终构建了 3 万句对左右的评测数据集。

6.3 数据集构建

这一节讲述了主宾互易数据集的构建，包括研究对象的建立，收集流程，以及最终数据集的相关统计量。

6.3.1 研究对象

假设 $\mathcal{S}$ 代表由所有语法表达正确、语义明确的简单小句组成的空间，并且该集合的句子成员 $l \in \mathcal{S}$，仅由主语 S，宾语 O，以及连接它们的动词 V 构成。这些在句子 s 中的成员可以表示为 s^S, s^O 和 s^V，该句子简记为 $s = SVO$，并且对应于该句子形式 s 的语义可以表示为：$m(s)$。我们同时定义一个主宾语互易的操作 s'，即 $s' = OVS$。通过该互易操作，可以将整个句子空间 $\mathcal{S}$ 划分为三个子空间，分别记作：I, E 和 U。它们的定义如下：

$$\begin{aligned} \mathcal{I} &= \{s \in \mathcal{S} | m(s) = m(s')\} \\ \mathcal{E} &= \{s \in \mathcal{S} | m(s) \neq m(s') \text{且} s' \in \mathcal{S}\} \\ \mathcal{U} &= \{s \in \mathcal{S} | s' \notin \mathcal{S}\} \end{aligned} \tag{10}$$

也就是说，I 中句子的主宾相互交换而语义不变（对应图10中的 A 类）,E 中的主宾交换后语义合规但相反（对应图10中的 B 类）,U 中句子主宾交换后语义不合规（图10中的 C 类）。值得注意的是，互易操作对于 $\mathcal{I}$ 和 $\mathcal{E}$ 两个空间具有对称性，即这两个空间中的任何一个元素在进行了可逆操作之后，它仍然处于原空间。用公式表达为：

$$\{s' | s \in \{\mathcal{I}, \mathcal{U}\}\} = \{\mathcal{I}, \mathcal{U}\}\} \tag{11}$$

而对于 $\mathcal{U}$ 空间中的句子，这种对称性没有得到保持。

之后我们收集相应的句子来构建这三个子空间。

6.3.2 收集流程

数据收集可以有两种方式：自上而下和自下而上。自上而下通过预先收集大规模的合法的原始语料句，再通过条件来过滤得到 $\mathcal{S}$，再设计规则分流出三类子空间 $\mathcal{I}$，$\mathcal{E}$ 和 $\mathcal{U}$。自下而上的方式则先为不同的子空间来挑选三个适合的组成成分 S、V、O，再将其组装成一个完整句子。自上而下的方式可以预先获取丰富的语句，且都是自然出现的真实语句，然而其过滤规则设计复杂；自下而上的方式则更容易确定空间中的成员，然而产生的句子都更加简单和单一。考虑到本文主要研究不常出现在语料中的简单句以及选取规则的可行性，本文主要采取自下而上的方式构建数据集。

（1）确定典型成员 对于每一类子空间，我们先确定其中的典型性成员，并且主要是通过动词的词汇语义进行确定的。对于互易后语义不变的集合 $\mathcal{I}$，我们借助以往的工

表 15 主宾互易后语义不变的语义类型举例说明

语义类型	举例
混合义	两份水泥配一份沙子 $\Longleftrightarrow$ 一份沙子配两份水泥
依附义	名字签空格里 $\Longleftrightarrow$ 空格里签名字
供给义	一间屋子住五个人 $\Longleftrightarrow$ 五个人住一间屋子
笼罩义	大楼笼罩着晨雾 $\Longleftrightarrow$ 晨雾笼罩着大楼
充满义	天空布满了乌云 $\Longleftrightarrow$ 乌云布满了天空
进入义	暗房透进一线光 $\Longleftrightarrow$ 一线光透进暗房

作[36], 主要从以下六类动词中进行选取：混合义、依附义、供给义、笼罩义、充满义和进入义。每一个意义下都有一些代表例句作为 $\mathcal{I}$ 中的典型成员。表 15列举了各个意义下典型的句子。对于 $\mathcal{E}$ 和 $\mathcal{U}$，本文借助一个中文动词语义网的词典资源[43]。该词典资源将常见动词分为分层的语义框架，包含动词源框架、初级框架、基本框架和微框架。例如源框架中包含“沟通”、“认知”、“感知”、“情绪”等。一般来说，同一初级框架下的动词处于同一个子空间中。最终我们为主宾互易后语义相反的集合 $\mathcal{E}$ 选取的源框架有“致使-动作”、“认知”、“伤害”、“感知”、“身体接触”等八大类；为互易后语义不合规的集合 $\mathcal{I}$ 选取了“动作”、“致使-位置”、“致使-行为”、“创造”、“存在”等 14 类。

再确认完动词后，便可以近一步缩小动词前后典型论元的范围。我们考虑如下的论元类别：专有名词、普通名词（个体、集体、物质、抽象）、时间、处所、方位、数量短语、代词和区别词。我们从考虑每一个论元类别的一些典型名词实例，以符合每个子空间的定义要求。

（2）数据增广 再确定了一些典型句子后，我们对于 S、V、O 各个部分的成分进行拓展。我们借助语义词典知网[22] 中的词汇语义相似度函数，寻找已确定的典型词汇成员的相似词汇，来丰富我们的数据。并且人工进行审核哪些词汇可以构成符合标准的句子。

（3）相关统计 最终数据的规模如表 16。

7 进度安排和预期成果

本项目研究计划从博士二年级开始一直持续到四年级中期，以下是分段的安排：

预期成果方面，计划在每一个主要的研究部分都可以通过实验和理论完整论证假设，并撰写论文，将其投稿并发表至领域内顶级的会议和期刊中。目前的前两个工作都已经被领域内的顶会 ACL 录用发表。

[43] http://mega.lt.cityu.edu.hk/~yufechen/#/

表 16 不同类型数据对应的统计量，$\mathcal{I}$ 中句子的主宾相互交换而语义不变（图10中 A 类）,$\mathcal{E}$ 中的主宾交换后语义合规但相反（图10中 B 类）,$\mathcal{U}$ 中句子主宾交换后语义不合规（图10中 C 类）

统计量	$\mathcal{I}$	$\mathcal{E}$	$\mathcal{U}$
句子数量	7817	13237	18056
主语数量	213	30	86
宾语数量	185	34	299
动词数量	93	126	352

表 17 进度安排

研究目标	时间				
	2023	2024 上	2024 下	2025 上	2025 下
文献调研、整理和归纳	✓	✓	✓	✓	
实验设计一：多义性中不确定性的建模	✓				
结果分析一	✓				
写作与发表	✓				
实验设计二：模型中多义性的反映		✓			
结果分析二		✓			
写作与发表		✓			
实验设计三：主宾互易数据集收集及评测			✓	✓	
结果分析三			✓	✓	
写作与发表			✓	✓	
论文修改和写作				✓	✓

主要参考文献

[1] KENTON J D M W C, TOUTANOVA L K. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of NAACL-HLT. [S.l.: s.n.], 2019: 4171-4186.

[2] OpenAI. Gpt-4 technical report[J/OL]. arXiv preprint arXiv:2303.08774, 2023. https://arxiv.org/abs/2303.08774.

[3] IYER V, CHEN P, BIRCH A. Towards effective disambiguation for machine translation with large language models[C]//WMT. [S.l.]: Association for Computational Linguistics, 2023: 482-495.

[4] MOSLEM Y, HAQUE R, WAY A. Adaptive machine translation with large language models[C]//Proceedings of the 2023 European Association for Machine Translation Conference (EAMT). [S.l.: s.n.], 2023.

[5] CHENG D, HUANG S, WEI F. Adapting large language models via reading comprehension[C/OL]//The Twelfth International Conference on Learning Representations. 2024. https://openreview.net/forum?id=y886UXPEZ0.

[6] MAO R, CHEN G, ZHANG X, et al. GPTEval: A Survey on Assessments of ChatGPT and GPT-4[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). [S.l.]: ELRA and ICCL, 2024: 7844-7866.

[7] BOMMASANI R, HUDSON D A, ADELI E, et al. On the opportunities and risks of foundation models[J]. CoRR, 2021, abs/2108.07258.

[8] 朱德熙. 语法讲义[M]. 北京: 商务印书馆, 1982.

[9] STERNEFELD W, ZIMMERMANN T E. Introduction to semantics: An essential guide to the composition of meaning (mouton textbook)[M]. [S.l.]: De Gruyter Mouton, 2013.

[10] SCHMITZ K. Economy in the lexicon: The role of polysemy[J]. 2022.

[11] NERLICH B, TODD Z, HERMAN V, et al. Polysemy: Flexible patterns of meaning in mind and language: volume 142[M]. [S.l.]: Walter de Gruyter, 2011.

[12] 张志毅, 张庆云. 词汇语义学[M]. 北京: 商务印书馆, 2012.

[13] DOWTY D R. Thematic proto-roles and argument selection[J]. Language, 1991, 67(3): 547-619.

[14] 李小凡, 张敏, 郭锐. 汉语多功能语法形式的语义地图研究[M]. 北京: 商务印书馆, 2015.

[15] HARRIS Z S. Distributional structure[J/OL]. Word, 1954, 10(23): 146-162. DOI: 10.1080/00437956.1954.11659520.

[16] FIRTH J R. A synopsis of linguistic theory 1930-1955[M]//Studies in Linguistic Analysis. Oxford: Blackwell, 1957: 1-32.

[17] FROMKIN V, RODMAN R, HYAMS N. An introduction to language (w/mla9e updates) [M]. [S.l.]: Cengage Learning, 2018.

[18] SHIELDS C. Order in multiplicity: Homonymy in the philosophy of aristotle[M]. [S.l.]: Oxford University Press, 2002.

[19] LYONS J. Semantics[M]. Cambridge: Cambridge University Press, 1977.

[20] CRUSE A. Meaning in language: An introduction to semantics and pragmatics[J]. 2004.

[21] MILLER G A, BECKWITH R, FELLBAUM C, et al. Introduction to wordnet: An online lexical database[J]. International journal of lexicography, 1990, 3(4): 235-244.

[22] DONG Z, DONG Q. Hownet-a hybrid language and knowledge resource[C]// International conference on natural language processing and knowledge engineering, 2003. Proceedings. 2003. [S.l.]: IEEE, 2003: 820-824.

[23] 沈家煊. 语法六讲[M]. 北京: 商务印书馆, 2011.

[24] 沈家煊. 名词和动词[M]. 北京: 商务印书馆, 2016.

[25] ROSCH E. Natural categories[J]. Cognitive Psychology, 1973, 4: 328-350.

[26] ROSCH E. Cognitive representations of semantic categories[J]. Journal of Experimental Psychology: General, 1975, 104: 192-233.

[27] MÀRQUEZ L, CARRERAS X, LITKOWSKI K C, et al. Special issue introduction: Semantic role labeling: An introduction to the special issue[J]. Computational Linguistics, 2008, 34(2): 145-159.

[28] FILLMORE C J. The case for case[M]. New York: Holt, Rinehart, and Winston, 1968: 1-88.

[29] JACKENDOFF R. Toward an explanatory semantic representation[J]. Linguistic Inquiry, 1976, 7(1): 89-150.

[30] 李临定. 施事、受事和句法分析[J]. 语文研究, 1984(04): 8-17.

[31] 张伯江. 从施受关系到句式语义[M]. 上海: 学林出版社, 2016.

[32] 马庆株. 自主动词与非自主动词[J]. 中国语言学报, 1988(3).

[33] 沈家煊. 不对称和标记论[M]. 南昌: 江西教育出版社, 1999.

[34] 冯国丽, 于秀金. 类型学视角下汉语的格配置与语义角色连续统[J/OL]. 外国语言文学, 2019, 36(05): 465-484. DOI: 10.19716/j.1672-4720.2019.05.02feng.

[35] 陈平. 试论汉语中三种句子成分与语义成分的配位原则[J]. 中国语文, 1994(03): 161-168.

[36] 李敏. 现代汉语主宾可互易句的考察[J]. 语言教学与研究, 1998.

[37] ZHANG W, DENG Y, LIU B Q, et al. Sentiment analysis in the era of large language models: A reality check[J/OL]. ArXiv, 2023, abs/2305.15005. https://api.semanticscholar.org/CorpusID:258866189.

[38] FINCH S E, PAEK E S, CHOI J D. Leveraging large language models for automated dialogue analysis[C]//Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue. [S.l.]: Association for Computational Linguistics, 2023: 202-215.

[39] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2009: 248-255.

[40] ZHU K, WANG J, ZHOU J, et al. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts[J]. arXiv preprint arXiv:2306.04528, 2023.

[41] WANG A, SINGH A, MICHAEL J, et al. Glue: A multi-task benchmark and analysis platform for natural language understanding[C]//Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. [S.l.]: Association for Computational Linguistics, 2018: 353-355.

[42] WANG A, PRUKSACHATKUN Y, NANGIA N, et al. Superglue: A stickier benchmark for general-purpose language understanding systems[C]//Advances in Neural Information Processing Systems 32. [S.l.: s.n.], 2019.

[43] HENDRYCKS D, BURNS C, BASART S, et al. Measuring massive multitask language understanding[C]//Proceedings of the International Conference on Learning Representations (ICLR). [S.l.: s.n.], 2020.

[44] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models[Z]. [S.l.: s.n.], 2020.

[45] ZHANG Y, TENG Z. Natural language processing: a machine learning perspective[M]. [S.l.]: Cambridge University Press, 2021.

[46] TSAI H, RIESA J, JOHNSON M, et al. Small and practical bert models for sequence labeling[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). [S.l.: s.n.], 2019: 3632-3636.

[47] LIU N F, GARDNER M, BELINKOV Y, et al. Linguistic knowledge and transferability of contextual representations[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 1073-1094.

[48] NAVIGLI R. Word sense disambiguation: A survey[J]. ACM computing surveys (CSUR), 2009, 41(2): 1-69.

[49] WU Z, CHEN Y, KAO B, et al. Perturbed masking: Parameter-free probing for analyzing and interpreting bert[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 4166-4176.

[50] GOLDBERG Y. Assessing bert's syntactic abilities[J]. arXiv preprint arXiv:1901.05287, 2019.

[51] JIN D, JIN Z, ZHOU J T, et al. Is bert really robust? a strong baseline for natural language attack on text classification and entailment[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.: s.n.], 2020.

[52] WANG L, LYU C, JI T, et al. Document-level machine translation with large language models[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. [S.l.]: Association for Computational Linguistics, 2023: 16646-16661.

[53] WEAVER W. Translation[C]//Proceedings of the Conference on Mechanical Translation. [S.l.: s.n.], 1952.

[54] EDMONDS P, COTTON S. Senseval-2: overview[C]//Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems. [S.l.: s.n.], 2001: 1-5.

[55] SNYDER B, PALMER M. The english all-words task[C]//Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text. [S.l.: s.n.], 2004: 41-43.

[56] PRADHAN S, LOPER E, DLIGACH D, et al. Semeval-2007 task-17: English lexical sample, srl and all words[C]//Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007). [S.l.: s.n.], 2007: 87-92.

[57] NAVIGLI R, JURGENS D, VANNELLA D. Semeval-2013 task 12: Multilingual word sense disambiguation[C]//Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). [S.l.: s.n.], 2013: 222-231.

[58] MORO A, NAVIGLI R. Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking[C]//Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015). [S.l.: s.n.], 2015: 288-297.

[59] MEI Q, XIE Y, YUAN W, et al. A turing test of whether ai chatbots are behaviorally similar to humans[J/OL]. Proceedings of the National Academy of Sciences, 2024, 121 (9): e2313925121. https://www.pnas.org/doi/abs/10.1073/pnas.2313925121.

[60] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. nature, 2015, 521(7553): 436-444.

[61] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(56): 1929-1958.

[62] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//Proceedings of the 32nd International Conference on Machine Learning. [S.l.]: PMLR, 2015: 448-456.

[63] SHORTEN C, KHOSHGOFTAAR T M. A survey on image data augmentation for deep learning[J]. J. Big Data, 2019, 6: 60.

[64] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//NeurIPS. [S.l.: s.n.], 2022.

[65] YU P, WANG T, GOLOVNEVA O, et al. ALERT: adapting language models to reasoning tasks[J]. CoRR, 2022, abs/2212.08286.

[66] CHEN Z, GAO Q. Curriculum: A broad-coverage benchmark for linguistic phenomena in natural language understanding[J/OL]. arXiv preprint arXiv:2204.06283, 2022. https://arxiv.org/abs/2204.06283.

[67] JI J, QIU T, CHEN B, et al. Ai alignment: A comprehensive survey[Z]. [S.l.: s.n.], 2024.

[68] BUOLAMWINI J, GEBRU T. Gender shades: Intersectional accuracy disparities in commercial gender classification[C]//Proceedings of the Conference on Fairness, Accountability, and Transparency. [S.l.]: PMLR, 2018: 77-91.

[69] NOBLE S U. Algorithms of oppression[M]. [S.l.]: New York University Press, 2018.

[70] ZHANG B H, LEMOINE B, MITCHELL M. Mitigating unwanted biases with adversarial learning[C]//Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. [S.l.: s.n.], 2018: 335-340.

[71] PAN A, CHAN J S, ZOU A, et al. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark[C]//Proceedings of the International Conference on Machine Learning (ICML). [S.l.: s.n.], 2023.

[72] HENDRYCKS D, BURNS C, BASART S, et al. Aligning ai with shared human values [C]//Proceedings of the International Conference on Learning Representations (ICLR). [S.l.: s.n.], 2021.

[73] Collective Intelligence Project. Introducing the collective intelligence project[EB/OL]. 2023. https://cip.org/whitepaper.

[74] HAGENDORFF T. The ethics of ai ethics: An evaluation of guidelines[J]. Minds and Machines, 2020, 30(1): 99-120.

[75] PANKOWSKA P K. Framework on ethical aspects of artificial intelligence, robotics and related technologies[R]. [S.l.]: European Parliament, 2020.

[76] ZOU A, PHAN L, CHEN S, et al. Representation engineering: A top-down approach to ai transparency[J]. arXiv preprint arXiv:2310.01405, 2023.

[77] BERESKA L, GAVVES E. Mechanistic interpretability for ai safety–a review[J]. arXiv preprint arXiv:2404.14082, 2024.

[78] WARSTADT A, PARRISH A, LIU H, et al. Blimp: The benchmark of linguistic minimal pairs for english[J]. Transactions of the Association for Computational Linguistics, 2020.

[79] CASALICCHIO G, MOLNAR C, BISCHL B. Visualizing the feature importance for black box models[C]//Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD). [S.l.: s.n.], 2018.

[80] SHAPLEY L S. A value for n-person games[M]. [S.l.]: Cambridge University Press, 1988.

[81] RIBEIRO M T, SINGH S, GUESTRIN C. "why should i trust you?": Explaining the predictions of any classifier[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (NAACL). [S.l.: s.n.], 2016.

[82] COVERT I C, LUNDBERG S, LEE S I. Explaining by removing: A unified framework for model explanation[J]. Journal of Machine Learning Research, 2021.

[83] JUMELET J. Evaluating and interpreting language models[Z]. [S.l.: s.n.], 2023.

[84] MIKOLOV T, CHEN K, CORRADO G, et al. Word2vec: Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[85] NAVIGLI R. Meaningful clustering of senses helps boost word sense disambiguation performance[C]//Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2006: 105-112.

[86] YAMAGIWA H, OYAMA M, SHIMODAIRA H. Discovering universal geometry in embeddings with ICA[C/OL]//BOUAMOR H, PINO J, BALI K. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023: 4647-4675. https://aclanthology.org/2023.emnlp-main.283. DOI: 10.18653/v1/2023.emnlp-main.283.

[87] LYU Q, APIDIANAKI M, CALLISON-BURCH C. Representation of lexical stylistic features in language models' embedding space[C/OL]//PALMER A, CAMACHO-COLLADOS J. Proceedings of the 12th Joint Conference on Lexical and Computational

Semantics (*SEM 2023). Toronto, Canada: Association for Computational Linguistics, 2023: 370-387. https://aclanthology.org/2023.starsem-1.32. DOI: 10.18653/v1/2023.starsem-1.32.

[88] ETTINGER A. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 34-48.

[89] HEWITT J, MANNING C D. A structural probe for finding syntax in word representations[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). [S.l.: s.n.], 2019: 4129-4138.

[90] TENNEY I, DAS D, PAVLICK E. Bert rediscovers the classical nlp pipeline[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2019: 4593-4601.

[91] ELHAGE N, NANDA N, OLSSON C, et al. A mathematical framework for transformer circuits[J]. Transformer Circuits Thread, 2021, 1: 1.

[92] OLSSON C, ELHAGE N, NANDA N, et al. In-context learning and induction heads[J]. arXiv e-prints, 2022: arXiv-2209.

[93] WANG K R, VARIENGIEN A, CONMY A, et al. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small[C]//NeurIPS ML Safety Workshop. [S.l.: s.n.], 2022.

[94] ELHAGE N, HUME T, OLSSON C, et al. Toymodelsof superposition[J]. 2022.

[95] 罗竹风, 等. 汉语大词典[M]. [出版地不详]: 汉语大词典出版社, 1986.

[96] AGIRRE E, DE LACALLE O L. Clustering wordnet word senses[J]. Recent Advances in Natural Language Processing III: Selected papers from RANLP, 2003: 121-130.

[97] ZADEH L A. Fuzzy sets[J]. Information and Control, 1965, 8: 338-353.

[98] GAWLIKOWSKI J, TASSI C R N, ALI M, et al. A survey of uncertainty in deep neural networks[J]. CoRR, 2021, abs/2107.03342.

[99] GAL Y, GHAHRAMANI Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning[C]//Proceedings of the 33rd International Conference on Machine Learning (ICML). [S.l.: s.n.], 2016: 1050-1059.

[100] NEAL R M. Bayesian learning for neural networks: volume 118[M]. [S.l.]: Springer Science & Business Media, 2012.

[101] MACKAY D J C. A practical bayesian framework for backpropagation networks[J]. Neural Computation, 1992, 4(3): 448-472.

[102] LAKSHMINARAYANAN B, PRITZEL A, BLUNDELL C. Simple and scalable predictive uncertainty estimation using deep ensembles[C]//Advances in Neural Information Processing Systems (NeurIPS). [S.l.: s.n.], 2017: 6402-6413.

[103] LIN Z, TRIVEDI S, SUN J. Generating with confidence: Uncertainty quantification for black-box large language models[J]. arXiv preprint arXiv:2305.19187, 2023.

[104] FARQUHAR S, KOSSEN J, KUHN L, et al. Detecting hallucinations in large language models using semantic entropy[J]. Nature, 2024, 630: 625-630.

[105] KOSSEN J, HAN J, RAZZAK M, et al. Semantic entropy probes: Robust and cheap hallucination detection in llms[J]. arXiv preprint arXiv:2406.15927, 2024.

[106] MILLER G A, CHODOROW M, LANDES S, et al. Using a semantic concordance for sense identification[C]//Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994. [S.l.: s.n.], 1994.

[107] FRANCIS W N, KUCERA H. Brown corpus manual[J]. Letters to the Editor, 1979, 5 (2): 7.

[108] MILLER G A, LEACOCK C, TENGI R, et al. A semantic concordance[C]//Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993. [S.l.: s.n.], 1993.

[109] TAGHIPOUR K, NG H T. One million sense-tagged instances for word sense disambiguation and induction[C]//Proceedings of the nineteenth conference on computational natural language learning. [S.l.: s.n.], 2015: 338-344.

[110] EISELE A, CHEN Y. Multiun: A multilingual corpus from united nation documents. [C]//LREC. [S.l.: s.n.], 2010.

[111] OCH F J, NEY H. Improved statistical alignment models[C]//Proceedings of the 38th annual meeting of the association for computational linguistics. [S.l.: s.n.], 2000: 440-447.

[112] RAGANATO A, CAMACHO-COLLADOS J, NAVIGLI R. Word sense disambiguation: A unified evaluation framework and empirical comparison[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. [S.l.: s.n.], 2017: 99-110.

[113] PETROLITO T, BOND F. A survey of wordnet annotated corpora[C]//Proceedings of the Seventh Global WordNet Conference. [S.l.: s.n.], 2014: 236-245.

[114] NAVIGLI R, PONZETTO S P. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network[J]. Artificial intelligence, 2012, 193: 217-250.

[115] LUO F, LIU T, XIA Q, et al. Incorporating glosses into neural word sense disambiguation[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2018: 2473-2482.

[116] HUANG L, SUN C, QIU X, et al. Glossbert: Bert for word sense disambiguation with gloss knowledge[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). [S.l.: s.n.], 2019: 3509-3514.

[117] KUMAR S, JAT S, SAXENA K, et al. Zero-shot word sense disambiguation using sense definition embeddings[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2019: 5670-5681.

[118] BEVILACQUA M, NAVIGLI R. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2020: 2854-2864.

[119] CONIA S, NAVIGLI R. Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. [S.l.: s.n.], 2021: 3269-3275.

[120] ZHANG X, ZHANG R, LI X, et al. Word sense disambiguation by refining target word embedding[C]//Proceedings of the ACM Web Conference 2023. [S.l.: s.n.], 2023: 1405-1414.

[121] BLEVINS T, ZETTLEMOYER L. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2020: 1006-1017.

[122] SU Y, ZHANG H, SONG Y, et al. Rare and zero-shot word sense disambiguation using z-reweighting[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2022: 4713-4723.

[123] WANG M, WANG Y. Word sense disambiguation: Towards interactive context exploitation from both word and sense perspectives[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). [S.l.: s.n.], 2021: 5218-5229.

[124] SCARLINI B, PASINI T, NAVIGLI R. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.: s.n.], 2020: 3528-3539.

[125] BARBA E, PASINI T, NAVIGLI R. Esc: Redesigning wsd with extractive sense comprehension[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S.l.: s.n.], 2021: 4661-4672.

[126] BARBA E, PROCOPIO L, NAVIGLI R. Consec: Word sense disambiguation as continuous sense comprehension[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 2021: 1492-1503.

[127] ZHANG G, LU W, PENG X, et al. Word sense disambiguation with knowledge-enhanced and local self-attention-based extractive sense comprehension[C]//Proceedings of the 29th International Conference on Computational Linguistics. [S.l.: s.n.], 2022: 4061-4070.

[128] TSENG Y H, KU M C, CHEN W L, et al. Vec2gloss: definition modeling leveraging contextualized vectors with wordnet gloss[J]. arXiv preprint arXiv:2305.17855, 2023.

[129] BEVILACQUA M, MARU M, NAVIGLI R. Generationary or "how we went beyond word sense inventories and learned to gloss" [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.: s.n.], 2020: 7207-7221.

[130] LI L, ROTH B, SPORLEDER C. Topic models for word sense disambiguation and token-based idiom detection[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2010: 1138-1147.

[131] VAN DE CRUYS T, APIDIANAKI M. Latent semantic word sense induction and disambiguation[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. [S.l.: s.n.], 2011: 1476-1485.

[132] LESK M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone[C]//Proceedings of the 5th annual international conference on Systems documentation. [S.l.: s.n.], 1986: 24-26.

[133] BANERJEE S, PEDERSEN T, et al. Extended gloss overlaps as a measure of semantic relatedness[C]//Ijcai: volume 3. [S.l.: s.n.], 2003: 805-810.

[134] WANG M, WANG Y. A synset relation-enhanced framework with a try-again mechanism for word sense disambiguation[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.: s.n.], 2020: 6229-6240.

[135] AGIRRE E, LÓPEZ DE LACALLE O, SOROA A. Random walks for knowledge-based word sense disambiguation[J]. Computational Linguistics, 2014, 40(1): 57-84.

[136] MORO A, RAGANATO A, NAVIGLI R. Entity linking meets word sense disambiguation: a unified approach[J]. Transactions of the Association for Computational Linguistics, 2014, 2: 231-244.

[137] SCOZZAFAVA F, MARU M, BRIGNONE F, et al. Personalized pagerank with syntagmatic information for multilingual word sense disambiguation[C]//Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations. [S.l.: s.n.], 2020: 37-46.

[138] TRIPODI R, NAVIGLI R. Game theory meets embeddings: a unified framework for word sense disambiguation[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). [S.l.: s.n.], 2019: 88-99.

[139] RADA R, MILI H, BICKNELL E, et al. Development and application of a metric on semantic nets[J]. IEEE transactions on systems, man, and cybernetics, 1989, 19(1): 17-30.

[140] LEACOCK C, CHODOROW M. Combining local context and wordnet similarity for word sense identification[J]. WordNet: An electronic lexical database, 1998, 49(2): 265-283.

[141] MARU M, SCOZZAFAVA F, MARTELLI F, et al. Syntagnet: Challenging supervised word sense disambiguation with lexical-semantic combinations[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). [S.l.: s.n.], 2019: 3534-3540.

[142] HINDLE D, ROOTH M. Structural ambiguity and lexical relations[J]. Computational linguistics, 1993, 19(1): 103-120.

[143] MCCARTHY D, CARROLL J. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences[J]. Computational Linguistics, 2003, 29(4): 639-654.

[144] ABNEY S, LIGHT M. Hiding a semantic class hierarchy in a markov model[C]//In Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing. [S.l.]: Citeseer, 1998.

[145] CLARK S, WEIR D. Class-based probability estimation using a semantic hierarchy[J]. Computational Linguistics, 2002, 28(2): 187-206.

[146] CIARAMITA M. Explaining away ambiguity: Learning verb selectional preference with bayesian networks[C]//Proc. International Conference of Computational Linguistics (2000). [S.l.: s.n.], 2000.

[147] RIVEST R L. Learning decision lists[J]. Machine learning, 1987, 2: 229-246.

[148] BLACK E. An experiment in computational discrimination of english word senses[J]. IBM Journal of research and development, 1988, 32(2): 185-194.

[149] MOONEY R. Comparative experiments on disambiguation word senses: An illustration of the role of bias in machine learning[C]//Proc. Conference on Empirical Methods in Natural Language Processing. [S.l.: s.n.], 1996: 82-91.

[150] TSATSARONIS G, VAZIRGIANNIS M, ANDROUTSOPOULOS I. Word sense disambiguation with spreading activation networks generated from thesauri.[C]//IJCAI: volume 27. [S.l.: s.n.], 2007: 223-252.

[151] DECADT B, HOSTE V, DAELEMANS W, et al. Gambl, genetic algorithm optimization of memory-based wsd[C]//3rd International workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3); held in conjunction with the 42nd Annual meeting of the Association for Computational Linguistics (ACL 2004). [S.l.]: Association for Computational Linguistics, 2004: 108-112.

[152] LEE Y K, NG H T. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation[C]//Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002). [S.l.: s.n.], 2002: 41-48.

[153] PUSTEJOVSKY J. The generative lexicon[M]. [S.l.]: MIT press, 1998.

[154] NORASET T, LIANG C, BIRNBAUM L, et al. Definition modeling: Learning to define word embeddings in natural language[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 31. [S.l.: s.n.], 2017.

[155] LI J, BAO Y, HUANG S, et al. Explicit semantic decomposition for definition generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2020: 708-717.

[156] CHANG T Y, CHEN Y N. What does this word mean? explaining contextualized embeddings with natural language definition[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). [S.l.: s.n.], 2019: 6064-6070.

[157] REID M, MARRESE-TAYLOR E, MATSUO Y. Vcdm: Leveraging variational bi-encoding and deep contextualized word representations for improved definition modeling[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.: s.n.], 2020: 6331-6344.

[158] ZHU R, NORASET T, LIU A, et al. Multi-sense definition modeling using word sense decompositions[J]. arXiv preprint arXiv:1909.09483, 2019.

[159] KABIRI A, COOK P. Evaluating a multi-sense definition generation model for multiple languages[C]//Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23. [S.l.]: Springer, 2020: 153-161.

[160] NI K, WANG W Y. Learning to explain non-standard english words and phrases[C]// Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers). [S.l.: s.n.], 2017: 413-417.

[161] ZHENG H, DAI D, LI L, et al. Decompose, fuse and generate: A formation-informed method for chinese definition generation[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S.l.: s.n.], 2021: 5524-5531.

[162] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[J]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014: 1532-1543.

[163] MIHALCEA R. Co-training and self-training for word sense disambiguation[C]// Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004. [S.l.: s.n.], 2004: 33-40.

[164] YAROWSKY D. Unsupervised word sense disambiguation rivaling supervised methods [C]//33rd annual meeting of the association for computational linguistics. [S.l.: s.n.], 1995: 189-196.

[165] ZHU J, WANG H, YAO T, et al. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification[C]//Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). [S.l.: s.n.], 2008: 1137-1144.

[166] KOHLI H. Transfer learning and augmentation for word sense disambiguation[C]// Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43. [S.l.]: Springer, 2021: 303-311.

[167] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[168] MARU M, CONIA S, BEVILACQUA M, et al. Nibbling at the hard core of word sense disambiguation[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2022: 4724-4737.

[169] LIU Z, LIU Y. Ambiguity meets uncertainty: Investigating uncertainty estimation for word sense disambiguation[C]//Findings of the Association for Computational Linguistics: ACL 2023. Toronto, Canada: Association for Computational Linguistics, 2023: 3963-3977.

[170] ZHANG X, HAUER B, KONDRAK G. Improving hownet-based chinese word sense disambiguation with translations[C]//Findings of the Association for Computational Linguistics: EMNLP 2022. [S.l.: s.n.], 2022: 4530-4536.

[171] CHEN H, HE T, JI D, et al. An unsupervised approach to chinese word sense disambiguation based on hownet[C]//International Journal of Computational Linguistics & Chinese Language Processing, Volume 10, Number 4, December 2005: Special Issue on Selected Papers from CLSW-5. [S.l.: s.n.], 2005: 473-482.

[172] PAN X, WANG H, OKA T, et al. Zuo zhuan ancient chinese dataset for word sense disambiguation[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop. [S.l.: s.n.], 2022: 129-135.

[173] ZHENG H, LI L, DAI D, et al. Leveraging word-formation knowledge for chinese word sense disambiguation[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. [S.l.: s.n.], 2021: 918-923.

[174] LI W, LU Q, LI W. Integrating collocation features in chinese word sense disambiguation [C]//Proceedings of the Fourth Sighan Workshop on Chinese Language Processing. [S.l.: s.n.], 2005.

[175] FAN C, LI Y. Chinese word sense disambiguation based on classification[C]//Intelligent Computing Theories and Application: 17th International Conference, ICIC 2021, Shenzhen, China, August 12–15, 2021, Proceedings, Part II. [S.l.]: Springer, 2021: 442-453.

[176] PILEHVAR M T, CAMACHO-COLLADOS J. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations[C]//Proceedings of NAACL-HLT. [S.l.: s.n.], 2019: 1267-1273.

[177] CALABRESE A, BEVILACQUA M, NAVIGLI R. Evilbert: Learning task-agnostic multimodal sense embeddings[C]//Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence. [S.l.: s.n.], 2021: 481-487.

[178] HERINGER H J, STRECKER B, WIMMER R. Syntax: Fragen, lösungen, alternativen [M]. [S.l.]: Fink, 1980.

[179] CAMPOLUNGO N, MARTELLI F, SAINA F, et al. Dibimt: A novel benchmark for measuring word sense disambiguation biases in machine translation[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2022: 4331-4352.

[180] GEIFMAN Y, EL-YANIV R. Selective classification for deep neural networks[J]. Advances in neural information processing systems, 2017, 30.

[181] VAZHENTSEV A, KUZMIN G, SHELMANOV A, et al. Uncertainty estimation of transformer predictions for misclassification detection[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2022: 8237-8252.

[182] GAL Y, ISLAM R, GHAHRAMANI Z. Deep bayesian active learning with image data [C]//International Conference on Machine Learning. [S.l.]: PMLR, 2017: 1183-1192.

[183] HOULSBY N, HUSZÁR F, GHAHRAMANI Z, et al. Bayesian active learning for classification and preference learning[J]. stat, 2011, 1050: 24.

[184] EL-YANIV R, et al. On the foundations of noise-free selective classification.[J]. Journal of Machine Learning Research, 2010, 11(5).

[185] XIN J, TANG R, YU Y, et al. The art of abstention: Selective prediction and error regularization for natural language processing[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). [S.l.: s.n.], 2021: 1040-1051.

[186] FOLK J R, MORRIS R K. Effects of syntactic category assignment on lexical ambiguity resolution in reading: An eye movement analysis[J]. Memory & Cognition, 2003, 31: 87-99.

[187] LIEBER R. Morphology and lexical semantics: volume 104[M]. [S.l.]: Cambridge University Press, 2004.

[188] LIU Z, KONG C, LIU Y, et al. Fantastic semantics and where to find them: Investigating which layers of generative llms reflect lexical semantics[Z]. [S.l.: s.n.], 2024.

[189] ETHAYARAJH K. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). [S.l.: s.n.], 2019: 55-65.

[190] JAWAHAR G, SAGOT B, SEDDAH D. What does BERT learn about the structure of language?[C/OL]//KORHONEN A, TRAUM D, MÀRQUEZ L. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 3651-3657. https://aclanthology.org/P19-1356. DOI: 10.18653/v1/P19-1356.

[191] WANG L, LI L, DAI D, et al. Label words are anchors: An information flow perspective for understanding in-context learning[C]//BOUAMOR H, PINO J, BALI K. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023.

[192] VOITA E, SENNRICH R, TITOV I. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives [C/OL]//INUI K, JIANG J, NG V, et al. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference

on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 4396-4406. https://aclanthology.org/D19-1448. DOI: 10.18653/v1/D19-1448.

[193] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: Open foundation and fine-tuned chat models[J]. arXiv preprint arXiv:2307.09288, 2023.

[194] JIANG T, HUANG S, LUAN Z, et al. Scaling sentence embeddings with large language models[J]. arXiv preprint arXiv:2307.16645, 2023.

[195] LOUREIRO D, JORGE A. Liaad at semdeep-5 challenge: Word-in-context (wic)[C]// Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5). [S.l.: s.n.], 2019: 1-5.

[196] MELAMUD O, GOLDBERGER J, DAGAN I. context2vec: Learning generic context embedding with bidirectional LSTM[C/OL]//RIEZLER S, GOLDBERG Y. Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. Berlin, Germany: Association for Computational Linguistics, 2016: 51-61. https://aclanthology.org/K16-1006. DOI: 10.18653/v1/K16-1006.

[197] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C/OL]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics, 2018: 2227-2237. https://aclanthology.org/N18-1202. DOI: 10.18653/v1/N18-1202.

[198] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models[Z]. [S.l.: s.n.], 2023.

[199] CHOMSKY N. Syntactic structures[M]. The Hague: Mouton, 1957.

[200] 龚千炎. 现代汉语里的受事主语句[J]. 中国语文, 1980(5).

[201] 张玲娟. 现代汉语主宾互易句研究[D]. [出版地不详]: 山东大学, 2012.

[202] 宋玉柱. 可逆句现代汉语特殊句式[M]. 南昌: 江西教育出版社, 1991.

[203] 任鹰. 主宾可换位供用句的语义条件分析[J]. 汉语学习, 1999.

[204] 李宇明. 存现结构中的主宾互易现象研究[M]. [出版地不详]: 商务印书馆, 2002.

[205] BOLEDA G. Distributional semantics and linguistic theory[J]. Annual Review of Linguistics, 2020, 6: 213-234.

[206] LI H, ABE N. Generalizing case frames using a thesaurus and the mdl principle[J]. Computational Linguistics, 24(2).

[207] GODDARD C, WIERZBICKA A. Words and meanings: Lexical semantics across domains, languages, and cultures[M]. [S.l.]: OUP Oxford, 2013.

[208] HASPELMATH M. Indefinite pronouns[M]. [S.l.]: Oxford University Press, 1997.

[209] 张敏. 如何从一个省的汉语方言语料导出人类语言共性规律: 湖南方言介词的语义地图研究[C]//上海师范大学语言研究所. 第五届汉语语法化问题国际学术讨论会论文集: 2009 年卷. 上海: 上海师范大学, 2009: 59-65.

[210] HASPELMATH M. Ditransitive constructions in the world’s languages[J]. Course handout for Leipzig Spring School on Linguistic Diversity. Leipzig, 2006.

[211] ZHANG Y. Semantic map approach to universals of conceptual correlations: a study on multifunctional repetitive grams[J]. Lingua Sinica, 2017, 3(1): 7.

[212] HASPELMATH M. The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison[M]//The new psychology of language. [S.l.]: Psychology Press, 2003: 211-242.

[213] CROFT W. Typology and universals[M/OL]. Cambridge: Cambridge University Press, 2003. DOI: 10.1017/CBO9780511840579.

[214] 马真. 现代汉语虚词研究方法论[M]. 修订本. 北京: 商务印书馆, 2016.

1 附录

1.1 各层的最优阈值

表 18 由 WiC 开发集所确定的 Llama2 模型在各层的最佳阈值

Layer Index	base	repeat	prompt
0	0.30	0.30	0.00
1	0.95	0.95	0.35
2	0.90	0.90	0.25
3	0.70	0.75	0.35
4	0.70	0.70	0.45
5	0.40	0.55	0.45
6	0.35	0.45	0.45
7	0.35	0.40	0.40
8	0.30	0.35	0.40
9	0.35	0.25	0.45
10	0.30	0.25	0.45
11	0.30	0.30	0.45
12	0.30	0.20	0.50
13	0.30	0.30	0.50
14	0.30	0.35	0.55
15	0.25	0.30	0.55
16	0.40	0.35	0.60
17	0.40	0.40	0.65
18	0.40	0.40	0.60
19	0.45	0.40	0.70
20	0.45	0.40	0.65
21	0.45	0.40	0.65
22	0.45	0.40	0.65
23	0.40	0.35	0.70
24	0.40	0.35	0.65
25	0.40	0.35	0.70
26	0.40	0.35	0.70
27	0.35	0.40	0.70
28	0.40	0.20	0.70
29	0.40	0.40	0.70
30	0.35	0.25	0.70
31	0.40	0.25	0.70
32	0.35	0.35	0.70