通过语言模型推断形容词上位词以增加开放英语 Wordnet 的连通性

Lorenzo Augello*
Università Cattolica del Sacro Cuore,
Milan, Italy.
lorenzo.augello01@icatt.it

John P. McCrae
Insight Research Ireland
Centre for Data Analytics,
University of Galway,
Galway, Ireland.
john@mccr.ae

Abstract

开放英语 Wordnet 是一个关键的资源，作为本体词汇-柠檬的一部分发布在语言链接开放数据云中。然而，该资源中缺失了许多链接，在本文中，我们研究如何在形容词之间建立上位关系。我们提出了关于上位关系的理论讨论，以及它在形容词中与名词和动词的不同之处。我们开发了一个新的形容词上位关系资源，并微调大型语言模型来预测形容词的上位关系，显示了 TaxoLLaMa 的方法可以适用于这个任务。

1 介绍

开放英文词网 (McCrae et al., 2019a, OEWN) 是普林斯顿词网 (Fellbaum, 1998) 的一个开源分支，它被建模并发布为 OntoLex 数据 (McCrae et al., 2017) 。这项工作旨在继续维护该资源的工作，并提供一个供链接数据资源连接的中心资源。然而，OEWN 本身并不是一个完全连接的图，如在动词层次结构上的最新研究所示 (McCrae, 2025) 。对于形容词和副词，有大量的同义词集 [1] 没有链接，这降低了作为语言链接开放数据 (LLOD) 云中一个中心资源的有效性 (McCrae et al., 2016) 。

最近大型语言模型的进步显著地重塑了词汇语义学领域 (Moskvoretskii et al., 2024b) 。这些模型在庞大语料库上进行训练，展示了推断细微词汇关系的新兴能力，并在分类提取和上位词发现任务中表现出色 (Bordea et al., 2016) 。我们的目标是利用这项能力来全面连接 OEWN 的图结构，然而，这些模型的性能主要是在名词和动词上进行了评估，关于它们是否能够有意义地捕捉形容词之间的复杂关系（如上位关系）的问题仍然未解，因为这在现有的数据集和词汇资源中代表性不足。

从大型文本语料库中已提取出了上位词，并在如 BLESS (Baroni and Lenci, 2011) 、EvaLution (Santus et al., 2015) 和 HyperLex (Vulić et al., 2017) 等被广泛使用的词汇关系分类基准中进行了建模。然而，其在形容词中的应用仍然显著缺乏探索，尽管它在理论和实践上都具有潜在的重要性。与名词和动词上位词不同的是，名词和动词的上位词通常依赖于明确定义的层次分类，而形容词之间的上位词本质上更为流动且依赖于上下文，其受多义性、可分级性和上下文模糊性等问题的影响，现有资源往往仅仅从其他关系的角度对这些问题进行调查 (Kennedy, 2007) ，(Murphy, 2003) 。

理解形容词之间的上位关系是至关重要的。为了探讨其中的问题，尽管其复杂性已被广泛承认，我们的目标是建立一个更清晰的形容词上位关系理论，以此提供对词汇意义更广泛组织的见解，并支持开发更具语义意识的语言模型。

在这项工作中，我们尝试解决形容词之间上位词关系的理论和实践空白，目标有两个。首先，我们提出一个解释形容词上位词关系的理论框架，并构建一个反映这种关系的高标准英语下位词-上位词对数据集，该数据集使用 OntoLex 模型以 RDF 格式发布。[2] 我们从现有的其他语言词汇资源开始，如波兰词网 (Maziarz et al., 2016) 和开放荷兰词网 (Postma et al., 2016) ，进行仔细的人为注释和验证，以解决语义歧义问题，并通过利用 OEWN 来确保英语数据集的可靠性。其次，我们探讨语言模型识别和解释形容词上位词关系的能力。我们评估它们在训练我们提出的高标准数据集之前和之后的表现，分析它们在多大程度上能够捕捉到这种复杂的词汇关系 [3] 。

检查模型在细粒度和主观语义关系上的表现可以探索和揭示其潜在的语言能力和偏见，这有助于正在进行的解释和改进其行为的努力。为了让模型更好地理解语言动态，我们认为开发专门的评估数据集是衡量和提高其分辨复杂语义现象能力的基础步骤 (Putra et al., 2024) 。

Work completed at Insight Research Ireland Centre for Data Analytics, University of Galway.

[1] “同义词集”或简称“synset”在 OntoLex 模型中等同于词汇概念

[2] https://github.com/lorenzoaugello/adjective-hypernymy

[3] https://huggingface.co/collections/loraug/hyper-discovery-68123c85fef0889c6559e674

本文的其余部分结构如下。在第 2 节中，我们讨论了关于上下位词关系、形容词语义和语言模型评估的相关工作，在第 3 节中，我们介绍了开放英文词网。第 4 节提出了我们的理论框架，而第 5 节描述了我们的形容词上下位词数据集的构建。第 6 节重点介绍了在此任务上对语言模型的评估，包括零样本测试和微调实验。第 7 节展示了结果和研究发现的总结。最后，第 8 节总结了本文并提出了未来研究的方向，同时阐述了本研究的局限性。

2 背景与相关工作

对于形容词的性质、功能、分类及其相互关系的理论化传统上始于对形容词如何修饰名词的分类，这体现了对于形容词定义的普遍共识，即形容词作为“与其结合的名词的修饰语”。然而，形容词的语义分类证明比名词和动词复杂得多，因为它们的行为常常难以被简单的本体论模型所捕捉。尽管进行了几次分类尝试，形容词仍然在词汇语义学中研究不足，并且缺乏一个普遍认可的理论框架。现有的分类通常停留在一般性质的高层次分类上，将形容词的同义集分组为几个被称为“超意义”的广泛语义类别，以创建分类法，正如 Dixon (1982) 和 Hundsnurscher and Splett (1982) 的首例所示。

将焦点转移到词汇蕴涵任务及其中特定关系——上位词关系时，我们面临更多的歧义。传统的上位词关系定义，比如提议一个术语 A 是另一个术语 B 的上位词，如果 A 的意义覆盖 B 的意义或者更广泛的 (Tjong Kim Sang, 2007)，应用于名词和动词时较为直接，因为它们的层级关系和结构更清晰。然而，形容词引入了与可分等级性、标度结构和上下文依赖相关的挑战 (Liu et al., 2023)。不像名词可以组织成链，每个下位词自然继承其上位词的属性 (Heyvaert, 2010)，形容词抵制简单的层级排列，更倾向于紧凑而不太垂直的标度。

值得注意的是，形容词的上位关系在基础词汇资源如普林斯顿 WordNet 中是不存在的。相反，形容词是通过反义关系（例如，湿–干）、语义相似性（例如，干–干旱）和归属关系（例如，犯罪–犯罪的）进行组织的，缺乏明确的层次关系建模，这可能导致形容词语义中的同义关系和上位关系的混淆。

现在的问题是关于形容词的上位词关系的存在。最近的工作表明，词汇关系不应被二元化地看待，而应被概念化为渐进的语义现象，其术语是根据类别成员原则沿连续尺度表示的 (Vulić et al., 2017)。然而，鉴于大多数现有的词汇资源和为词汇语义任务创建的数据集源于普林斯顿 WordNet 的结构——而后者并未纳入形容词的上位词关系——这一关系的表示和形式化仍然不够完善。

一些其他语言的词汇资源已经采取措施来模拟形容词的上位词关系。例如，GermaNet (Hamp and Feldweg, 1997) 放弃了普林斯顿 WordNet 的基于聚类的方法，而采用了类似于名词和动词的层次结构来组织形容词，例如，“gut”（“好”）是“toll”（“棒”）的上位词。在波兰 WordNet 中也可以观察到类似的层次映射，其中普林斯顿 WordNet 的原始“哑铃模型”已转变为一种纵向下位词结构 (Rudnicka et al., 2016)。在开放荷兰 WordNet 中，尽管反义关系仍然是形容词的主要关系，但也引入了层次上位词关系的实例，例如“knots-gek, stapelgek, krankjorum, knettergek”（“非常疯狂”）作为“gek, dwaas”（“疯狂”）的下位词 (Maks et al., 2008)。

在自然语言处理社区的背景下，关于上位词关系的研究已经在许多工作中进行了探索，从其在文本语料库中的自动提取到围绕语言模型的知识和分类能力的研究。已经进行了各种评估研究以考察语言模型的表现，但它们主要集中在名词、动词或一般自然语言推理任务上。诸如基于提示的上位词知识评估、数据集增强和基准测试策略等特定技术已经被提出，但仍然没有专门针对形容词上位词的研究。因此，虽然在理解其他词类的词汇蕴含方面取得了显著进展，但专门针对形容词的研究仍然是一个未被探索的领域。

3 开放英语词网

开放英文词网 (McCrae et al., 2019b) 是一个全面的、开源的英语词汇资源，源自原始的普林斯顿词网 (Fellbaum, 1998)，并使用 W3C OntoLex-Lemon 模型 (McCrae et al., 2017) 研发而成。作为 OntoLex 资源，OEWN 在网上发布词汇数据，符合关联数据原则，从而促进跨语言和语义资源的互操作性。

OEWN 在语言链接开放数据（LLOD）云中扮演着核心角色，(McCrae et al., 2016) 作为一个基础资源来表示词汇语义，促进跨语言比较，并与其他语言资源进行连接。它使用 OntoLex-Lemon 模型作为一种标准化的词汇来以 RDF 表示词汇信息。全球词网协会开发了用于词网表示的格式 (McCrae et al., 2021)，这些格式扩展了 OntoLex 的核心原则。这些格式支持 XML、JSON(-LD) 和 RDF/Turtle 中的多种序列化——提供与其他 OntoLex 资源的互操作性。这些格式已经被 OEWN 和开放多语言词网（OMW）中的其他词网采纳，以确保兼容性和可扩展性。

在 RDF 中，GWA 格式直接使用 OntoLex 类编码 `LexicalEntry`、`LexicalSense` 和 `LexicalConcept`，而 WordNet 特定的元数据和关系（例如，义项键、同义词集排序）则通过 https://globalwordnet.github.io/schemas/wn 处的专用本体进行建模。

4 理论建议

尽管形容词的上位关系很难定义且界限不清，但我们认为这是一个必要的关系，可以通过引入意义和用法的层次结构来帮助语言的语义组织，这个层次结构并不限于广泛类别的简单分类或模糊且常常混淆的水平和对立关系。与同义关系的区别实际上非常微妙，但我们认为在同一语义领域内存在一些意义比其他形容词更广泛的形容词，这些需要被区分（例如：认知 - 心理 - 非物质；令人不快 - 不愉快 - 消极）。

普林斯顿 WordNet 包含 117,659 个同义词集和 84,428 个上义词-下义词关系，但这些仅限于名词和动词，不包括形容词和副词。此外，目前没有可靠的数据集来表示英语中形容词的这种关系。面对高质量训练数据有限的情况，本研究的主要目标之一是创建一个可靠的初始数据集，其中包含从 OEWN 中提取的形容词对及其定义，以解决词义消歧问题。

在从现有词汇资源中提取下位词-上位词对的注释和可靠性评估过程中，遵循了替代原则（其他词汇替代的应用见 McCarthy and Navigli (2007)）：在语境中，上位词应该可以替代下位词，从而使所得句子在可接受的普遍层面上保留原意，且不矛盾或不需要强制解释。虽然语义拓展是预期且必要的，但下位词的意义需要在上位词的可能解释中被无歧义地包含。因此，上位词关系以意义的包含为特征，而非等同，区别于同义关系。在注释过程中，如果候选对的两个形容词之间的联系没有被注释者直观且直接地感知到，则该对被弃用，因为我们不希望出现可疑的推论或多种合理解读。鉴于上位词关系本身难以定义——其界限模糊且常常具有灵活和可争议的性质——因此，仅保留那些被认为最可靠的对（当然，毫无疑问，有些可能仍会引起争议，再次强调了这种关系的主观性）。

识别两个形容词之间的上位关系的主要挑战之一在于它们的多义性。第一个例子是 “cold”，它可以指温度的语义领域，也可以指人类行为和性格方面。另一个例子是 “hard”，一种意义上是指物理上坚固的材料，另一种则是指更抽象的困难概念。因此，形容词的意义及其关系不仅仅在单独的情况下被定义，更重要的是要根据它们所伴随的名词和它们唤起的语义领域来定义：“cold” 可以描述环境，也可以形容一个人；“hard” 可以指材料，也可以说明一个问题。在这方面，上下文起着至关重要的作用。虽然这对人类标注者来说已经是一个挑战，但对于必须处理语义消歧的语言模型来说，这变得更加复杂（见第 5 节）。此外，许多英语单词可以根据上下文和用法来承担不止一种词性，而不进行词形变化，因为英语的形态丰富性非常有限。第一个例子是单词 “clean”，它可以作为动词和形容词使用。一个更广泛且更常见的情况涉及现在分词和过去分词，有时被解释为动词，有时为形容词。这种歧义也可以扩展到名词（上面提到的单词 “cold” 可以有多种形容词意义，但也可以是名词：“涉及鼻子和呼吸道的轻度病毒感染”），鉴于这种现象在英语中的普遍性，有必要在构建黄金标准数据集时结合 OEWN 定义来包括下位词和上位词（也可参见 Moskvoretskii et al. (2024b) 的实验以了解定义为何至关重要）。

相比于更直接的名词上位关系，我们认为形容词的上位关系需要扎根于意义消歧，通过基于替代的包含测试进行操作，并在上下文中进行评估。我们的黄金标准数据集利用了这些原则，旨在提供初步但可靠的资源，为进一步研究形容词的语义及其在计算系统中的建模提供支持。

5 数据集创建

金标准数据集的构建基于已有的词汇资源，利用了显式编码形容词上位关系的语言的词网：开放荷兰词网（ODWN）和波兰词网（plWN）。选择这两个资源的原因是，与普林斯顿词网不同，它们以层级结构组织形容词，使其成为我们目的的合适起点。

通过使用 `wn` Python 库[4]，从这两种资源中自动提取了形容词的下位词-上位词对，并从总数中随机选择了 450 对作为初始集合（ODWN 取 166 对，plWN 取 284 对）。只考虑形容词，并包括了子类下位词和父类上位词节点。每个提取的对随后使用双语在线词典（维基词典和剑桥词典）手动翻译成英语，为了最大程度减少主观偏见和随意选择，我们一致选择了第一个建议的翻译。

之后，每对词条由两名注释者单独审查。如果该对的两个形容词出现在 OEWN 的同一个同义词集中——这表明它们是同义关系而不是层级关系——则该对词条会被舍弃，保持原来的关系：例如，“difficult” 和 “hard” 在 plWN 中是上位词和下位词关系，但没有被包括在内，因为它们在 OEWN 中是同义词（“不容易；需

[4] https://pypi.org/project/wn/

Annotation	hypo	hyper
yes-no	intelligent	rational
yes-maybe	limitless	vast
yes-yes	lucid	aware
no-maybe	multiple	plural
maybe-maybe	unneeded	useless
no-no	productive	rich

Table 1: 两个注释者对形容词对进行注释时的同意和分歧示例。

要很大的体力或智力努力才能完成、理解或忍受”）。当一个下位词有多个上位词时，如果它们属于同一个 OEWN 同义词集就保留，否则只选择最可靠的一个并由注释者达成一致，其他的则被舍弃。

因此，从最初的 450 对中，有 148 对被舍弃，最终的英语金标准包含 302 对形容词：其中 92 对来源于 ODWN，170 对来源于 plWN，还有 40 对来自两者之一（18 对来自 ODWN，22 对来自 plWN），但这 40 对中提出了一种更为可靠的替代上位词并获得注释者的认可（例如，plWN 中 “deft” 的上位词是"effective"，但建议并核准了 “skillful” 作为替代并保留在数据集中）。不同资源的接受率相似，ODWN 的保留率为 0.55（92/166），plWN 为 0.60（170/284）。此外，如第 4 节已经提到的那样，对于包含在金标准中的每个形容词，其相应的英文定义在从 OEWN 中检索之后也包括在内。这对于确保注释的关系反映形容词的预期意义而非可能的多义变体是至关重要的。

标注过程始于一个初始的小型数据集，其中包含 50 对形容词对，由两名标注者使用三标签分类系统（是，否，也许）进行标注。所有负面一致（否-否）、强烈分歧（是-否）、较弱分歧（否-也许）和共同疑虑（也许-也许）的案例被舍弃，而完全一致（是-是）的案例被保留，部分一致（是-也许）的案例则进一步讨论（例子见表 1）。从第一个样本中，选择了 62 对%（31 对形容词对）（Cohen’s kappa $\kappa = 0.65$）。从第二个样本的 100 对中，保留了 69 对%（$\kappa = 0.64$），从第三个样本的 300 对中，选择了 67 对%（201 对）（$\kappa = 0.61$）。然后引入第三名标注者对从第三个较大样本（300 对）中随机选择的 100 对进行标注，提供了额外的讨论和验证（Fleiss’ kappa $\kappa = 0.48$）。

鉴于语义关系任务引入高水平的主观性，其难度较大，因此标注者之间的协议程度属中等，但仍然是可以接受的、一致的，并且在不同的样本之间具有可比性。大多数分歧涉及 “可能” 标签，而不是直接矛盾（在第一个样本中有 15 次出现 “可能” 标签，第二个样本中有 24 次，第三个样本中有 93 次）。

除了每个下位词都与一个单一的准确上位词关联的主要黄金标准之外，后来还开发了数据集的第二个版本，用于模型微调和评估（参见章节 6.1）。在这里，每个上位词的同义词根据它们在 OEWN 中的同义词集成员资格被添加，以便解决存在多重语义上正确的上位词的情况（参见表格 2）。这样做的动机不仅在于给定形容词的上位词的自然多样性，还在于减少在模型评估过程中基于定义的提示的可能影响。

6 方法论

从形容词的上位词定义到创建基准数据集的问题，可以显而易见的是，词汇蕴涵这一任务已经在许多人类理解上呈现出许多问题。关于语言模型的能力，有一些尝试评估它们的可靠性，或者采取二元分类的方法（选择正确的上位词），或者采取生成的方法（给定一个下位词，预测其最可能的上位词）。

在本研究中，我们使用了两个模型：TaxoLLaMa 和 SmolLM-360M-Instruct。我们首先测试它们在上下位词发现任务中的能力，然后在我们的基准数据集上微调它们，以便探索它们从中捕捉语义知识的能力。

TaxoLLaMa (Moskvoretskii et al., 2024a) 是 LLaMA-2-7b 模型 (Touvron et al., 2023) 的微调版本，该模型在 WordNet 数据集上针对 16 个与分类法相关的任务进行了训练，并在不同领域和语言的上位词发现上达到了 SoTA 的效果。然而，由于它既未在形容词示例上进行训练，也未进行测试，因此在预测形容词时，其性能预计会较低。

SmolLM-360M-Instruct 是 SmolLM 的一部分，SmolLM 是一组最先进的小型模型 [5]。它通过在多个指令数据集上的监督微调来优化以执行指令任务。尽管它比 TaxoLLaMa 小得多，但其小巧的体积使其成为快速微调和评估专门任务（如形容词上义词发现）的理想选择，使我们能够探索在提供有限数据的情况下小型模型是否能获取语义关系。

6.1 训练细节

我们使用 Unsloth 方法进行微调，这使我们能够将模型量化为 4 位，并使用 LoRA 进行训练，从而在不影响性能的情况下减少内存和计算需求。这种方法特别适合使用我们的小型数据集，并使得像 TaxoLLaMa 这样的小型和大型模型在有限的硬件资源和易于访问的情况下进行微调。我们使用 SFTTrainer 类微调模型，进行了 60 次优化步骤的训练，学习率设置为 2e-4。每个设备的批量大小设置为 2，并使用 4 个步骤

[5] https://huggingface.co/blog/smollm

Dataset	hyponym-lemma	hypo_definition	hypernym-lemma	hyper_definition
single	relaxed	without strain or anxiety	calm	not agitated; without losing self-possession
multiple	relaxed	without strain or anxiety	calm, serene, tranquil, unagitated	not agitated; without losing self-possession

Table 2: 以下是从两个版本的黄金标准中选取的一对示例，首先显示输入下位词（“relaxed”）的一个精确上位词（“calm”），然后显示在 OEWN 中找到的其同义词。

进行梯度累积，模拟批量大小为 8。利用我们的金标准数据集，我们在由 211 个项目（占数据集 70%）构成的训练集上微调模型。下面是一个训练样本，包含人类用户提出的输入问题和 GPT 助手预期的输出答案，遵循聊天模板。

> （来自: human）“词语‘复杂’（定义：‘难以分析或理解’）的上位词是什么?”,
>
> （来自: gpt）“上位词是：‘困难，辛苦’（定义：‘不容易；需要很大的体力或脑力来完成、理解或忍受’）。”

训练在两个不同的设置下进行：首先，模型使用原始金标准数据集进行微调（以下简称为“单一”），在每个下义词与其相关的上义词之间保持一对一的对应关系；然后开发了一个不同的金标准（“多重”），在单一数据集中加上上义词的同义词，以实现每个下义词与其相关上义词之间的一对多对应关系（见表 2）。为了保持一致的标准并避免歧义，我们依靠 OEWN 并添加了所有属于原始单个上义词同义词集的形容词。这样做的原因主要有两个。首先，鉴于一个下义词可以有多于一个的上义词，只在金标准中包括一个正确的确切上义词可能会遗漏其他可能的候选项（例如，“合适的”在单一数据集中有“合适的”作为其确切上义词，但在多重数据集中增加了“适当的”和“适合的”，属于同一个同义词集：适合或为一个场合或用途适应而定）。其次，使用单一数据集训练的模型在测试时对每个下义词只会输出一个上义词，而如果用多个可能的上义词进行训练，它们将包含更多的预测并提高其语义知识。

6.2 提示

在零样本评估期间和微调之后，我们使用了两种不同的提示设置：首先，我们仅向模型提供输入的下义词形容词，然后我们也给它们提供下义词定义。对于零样本提示，我们遵循了原本为 TaxoLLaMa 使用的以下格式：

> <s>[INST] «SYS» 你是一个乐于助人的助手。列出所有可能的单词，用逗号分隔。你的回答不应该包含除逗号分隔的单词以外的任何内容 «/SYS»
>
> 下位词: 幽默的 (充满幽默或具有幽默特征的) | 上位词: [/INST]

然而，在微调之后我们遵循以下格式：

> messages = [‘‘来自”: ‘‘人类”, ‘‘值”: ‘‘下位词: ‘‘振奋人心的’’（定义: ‘‘赋予力量和活力’’）的上位词是什么？ ” ,]

7 结果

原始基础模型 TaxoLLaMa 达到了 SoTA 的结果，并在英语中的 MRR（平均倒数排名）得分为 54.39，但这仅是在 WordNet-3.0 图中采样的动词和名词上进行训练的。因此，其在形容词的上位词发现任务中的表现预计会较低。为了评估这一点，我们在一个零样本设置中使用我们的测试集（91 对，占总黄金标准的 30%）进行测试，记录到当模型没有被提示输入下位词的定义时，MRR 为 9.4，而当给出定义时则提高到 25.8。经过在包含同义词的多个数据集上进行微调后，这些得分分别改善至 23.6 和 33.3。

在基本的 TaxoLLaMa 中，处理形容词并将其与其他词性区分开是个主要的难题：在没有定义的情况下进行零样本测试时，58 % 的情况下仅输出名词，21 % 仅输出形容词，21 % 同时输出两者（通过提供定义，相应的数字变为 14、44 和 42 %）。在对单一数据集进行微调后，预测到的形容词数量显著增加：没有定义时为 95 %，有定义时为 100 %（在多个数据集上微调时，两者均为 100 %）。如表格 3 所示，这一改进是通过 TaxoLLaMa 和 SmolLM-360M-Instruct 共同实现的。鉴于词性识别和消歧的难度，我们认为这一结果几乎与正确的上位词预测同样重要和有意义。

微调之后，这些模型在两个设置下针对两个不同的数据集进行了测试，以先评估它们推断给定下位词的精确正确上位词的能力（针对单一数据集），然后引入在输出中给出同义词的可能性（针对多数据集）。

表 4 显示了在第一种设置下的结果，其中 TaxoLLaMa-ft-multi 在提示中不提供和提供定义的情况下都达到了最佳得分。有趣的是，提

Model	Setting	No def	With def
TaxoLLaMa	Zero-shot	0.39	0.78
TaxoLLaMa	ft-single	0.96	1.00
TaxoLLaMa	ft-multi	1.00	1.00
SmolLM-360M-Instr	Zero-shot	0.69	0.77
SmolLM-360M-Instr	ft-single	0.79	0.96
SmolLM-360M-Instr	ft-multi	1.00	1.00

Table 3: 在预测正确的词性（形容词）的 F1 分数表现中，分别在微调前（零样本）、使用单一数据集训练（ft-single）和使用多个数据集训练（ft-multi）情况下，不含定义和含定义的情形下的表现。

供定义并没有提高 TaxoLLaMa-ft-single 的性能。乍一看，这似乎令人惊讶。然而，我们需要考虑到这个模型只生成一个形容词作为输出，并且这种选择往往受到输入定义中词汇重叠的强烈影响：例如，“extant”（被定义为“仍然存在；未灭绝、未毁坏或未丢失”）的正确上位词是“real”，但 TaxoLLaMa-ft-single 因定义中的词汇而预测“existent”。

Model	Setting	No def	With def
TaxoLLaMa	Zero-shot	0.15	0.39
TaxoLLaMa	ft-single	0.32	0.31
TaxoLLaMa	ft-multi	0.35	0.44
SmolLM-360M-Instr	Zero-shot	0.13	0.16
SmolLM-360M-Instr	ft-single	0.14	0.15
SmolLM-360M-Instr	ft-multi	0.21	0.25

Table 4: 在单数据集上评估性能，以预测准确的正确上位词，在微调之前，在单数据集上训练和在多个数据集上训练时，无论是否使用定义。考虑到除 TaxoLLaMa 零样本模型外，所有其他模型都输出一个单一的上位词，因此精度和召回率是相等的，仅报告一个值。

由于定义导致的这种变异性和影响也是将同义词引入多个黄金标准的动机之一（在 OEWN 中，“真实”和“存在”属于同一同义词集）。通过允许同义词，我们可以认为模型预测出的语义上正确的上位词，即使不是我们预期的那个，也是可以接受的。此外，重要的是要注意，基础模型 TaxoLLaMa 训练的目标是输出上位词列表，而 TaxoLLaMa-ft-single 仅产生单一预测。这种结构差异增加了基础模型输出中包含正确上位词的几率（结果是更高的 F1 值，0.39 对比 0.31）。然而，当仅评估基础模型 TaxoLLaMa 输出列表中排名第一的上位词时，其性能下降，14 个正确的上位词中只有 7 个出现在首位。

表格 5 所示的结果是在针对多数据集的第二次评估设置中获得的，显示了 TaxoLLaMa-ft-multi 优于其他模型的表现。当在提示中包含定义时，我们观察到模型性能的一致改进，揭示了在消歧中加入定义的重要性。此外，与在单一数据集上训练的对应模型（TaxoLLaMa-ft-single 和 SmolLM-ft-single）相比，TaxoLLaMa-ft-multi 和 SmolLM-ft-multi 都表现出明显的提升，突显了在训练中允许多个有效上位词以捕捉上位关系微妙特性所带来的益处。

8 结论

在这篇论文中，我们从理论和计算角度探索了形容词上位词关系这一较少研究的关系。我们提出了一个基于语义包容性和语境可替代性的定义，从而将其与同义关系和其他关系区分开来。利用这一框架，我们通过调整波兰语、荷兰语及开放英语词网中储存的词汇信息，构建了一个包含两个版本的英语形容词上位词关系的黄金标准数据集。我们的数据集通过人工标注和基于同义词集的消歧进行验证，为未来的研究提供了一个可靠、小型且初步的基准。

然后，我们评估了语言模型——大型的 TaxoLLaMa 和较小的 SmolLM-360M-Instruct——在超级类别发现任务上的能力，分别在零样本和微调设置中。我们的结果表明，这些模型最初在形容词超级类别上表现得很困难，尤其是在词性歧义和语义多义性问题上。然而，在我们的数据集上进行微调之后，特别是在同义词增强变体上，它们的表现有所改善，突出了任务特定训练数据的价值。我们还发现，向模型提供输入形容词的明确定义能够提高其识别正确超级类别的能力。这强调了词义消歧在首先识别然后建模形容词意义方面的核心作用。

作为未来的工作，我们希望 a) 扩展金标准数据集，以达到更高的覆盖率和更广泛的通用性，从而在训练和评估语言模型时允许更好的理论描述和更高的可靠性，b) 在 OEWN 中建模形容词的上位关系，识别并表示该资源中所有形容词的上位关系，c) 支持 OntoLex 与 NLP 社区之间的互操作性，推动在下游应用中使用形容词的上位关系，d) 将语言模型的评估扩展到形容词之间的其他语义关系，以探索它们在理论上的差异以及它们在计算上被捕捉得如何。

9

限制

1. 数据集规模：黄金标准在规模上有限，这可能限制其理论的完整性（在建模形容词上位词的全谱方面），也限制其实用性（用于模型训练和评估，特别是在未见或歧义的形容词对上），从而限制了其泛化能力和词汇覆盖范围。

2. 模型评估：仅使用了两个语言模型，因此在必然排除其他现有架构、规模和训练数

Model	Setting	No def			With def		
		P	R	F-M	P	R	F-M
TaxoLLaMa	Zero-shot	0.04	0.13	0.06	0.07	0.23	0.11
TaxoLLaMa	ft-single	0.14	0.14	0.14	0.16	0.16	0.16
TaxoLLaMa	ft-multi	0.15	0.20	0.17	0.28	0.25	0.26
SmolLM-360M-Instr	Zero-shot	0.07	0.07	0.07	0.09	0.09	0.09
SmolLM-360M-Instr	ft-single	0.09	0.09	0.09	0.10	0.10	0.10
SmolLM-360M-Instr	ft-multi	0.16	0.10	0.12	0.20	0.14	0.16

Table 5: 在对多个数据集进行预测时，评估在微调之前、单个数据集训练时和多个数据集训练时，是否包含定义情况下预测可能上位词列表的性能。对于那些仅输出一个上位词的模型（如 TaxoLLama-ft-single、SmolLM-Zero-shot 和 SmolLM-ft-single），精确度、召回率和 F-度量值是相同的。

据多样性的情况下，这限制了关于模型能力的结论。

3. 标注的主观性：标注过程虽然基于精心定义的标准和多标注者的一致性，但它引入了一定程度的主观性，这种主观性因形容词上位词关系本身的分级、细微差别及多义性而更加突出。

4. 语言：数据集、理论和评估都仅限于英语。因此，所观察到的一些现象可能无法在跨语言上普遍化，因为英语形容词的语义行为可能与其他语言不同。

References

Marco Baroni and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10, Edinburgh, UK. Association for Computational Linguistics.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. SemEval-2016 task 13: Taxonomy extraction evaluation (TExEval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1081–1091, San Diego, California. Association for Computational Linguistics.

R. M. W. Dixon. 1982. *Where have All the Adjectives Gone?* De Gruyter Mouton, Berlin, New York.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Frans Heyvaert. 2010. An outline for a semantic categorization of adjectives. In *Proceedings of the 14th EURALEX International Congress*, pages 1309–1318, Leeuwarden/Ljouwert, The Netherlands. Fryske Akademy.

Franz Hundsnurscher and Jochen Splett. 1982. *Grundlegung Einer Semantischen Beschreibung der Adjektive des Deutschen*, pages 16–47. VS Verlag für Sozialwissenschaften, Wiesbaden.

Cristopher Kennedy. 2007. Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics & Philosophy*, 30:1–45.

Wei Liu, Ming Xiang, and Nai Ding. 2023. Adjective scale probe: Can language models encode formal semantics information? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13282–13290.

Isa Maks, Piek Vossen, Roxane Segers, and Hennie van der Vliet. 2008. Adjectives in the Dutch semantic lexical database CORNETTO. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plWordNet 3.0 – a comprehensive lexical-semantic resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2259–2268, Osaka, Japan. The COLING 2016 Organizing Committee.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic. Association for Computational Linguistics.

John P. McCrae. 2025. Renovating the verb hierarchy of english wordnet. In *Global WordNet Conference 2025*.

John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017*, pages 587–597.

John P. McCrae, Christian Chiarcos, Francis Bond, Philipp Cimiano, Thierry Declerck, Gerard de Melo, Jorge Gracia, Sebastian Hellmann, Bettina Klimek, Steven Moran, Petya Osenova, Antonio Pareja-Lora,

and Jonathan Pool. 2016. The open linguistics working group: Developing the linguistic linked open data cloud. In *10th Language Resource and Evaluation Conference (LREC)*, pages 2435–2441.

John P. McCrae, Michael Wayne Goodman, Francis Bond, Alexandre Rademaker, Ewa Rudnicka, and Luis Morgado Da Costa. 2021. The globalwordnet formats: Updates for 2020. In *Proceedings of the 11th Global Wordnet Conference*, pages 91–99.

John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. 2019a. English WordNet 2019 – an open-source WordNet for English. In *Proceedings of the 10th Global Wordnet Conference*, pages 245–252, Wroclaw, Poland. Global Wordnet Association.

John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. 2019b. English wordnet 2019 – an open-source wordnet for english. In *Proceedings of the 10th Global WordNet Conference –GWC 2019*.

Viktor Moskvoretskii, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, and Irina Nikishina. 2024a. TaxoLLaMA: WordNet-based model for solving multiple lexical semantic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 2331–2350. Association for Computational Linguistics.

Viktor Moskvoretskii, Alexander Panchenko, and Irina Nikishina. 2024b. Are large language models good at lexical semantics? a case of taxonomy learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1498–1510, Torino, Italia. ELRA and ICCL.

M. Lynne Murphy. 2003. Semantic relations and the lexicon: antonymy, synonymy, and other paradigms. *Acta Linguistica Hafniensia*, 36(1):185–189.

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016. Open Dutch WordNet. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 302–310, Bucharest, Romania. Global Wordnet Association.

I. Made Suwija Putra, Daniel Siahaan, and Ahmad Saikhu. 2024. Recognizing textual entailment: A review of resources, approaches, applications, and challenges. *ICT Express*, 10(1):132–155. Publisher Copyright: © 2023 The Author(s).

Ewa Rudnicka, Wojciech Witkowski, and Katarzyna Podlaska. 2016. Challenges of adjective mapping between plWordNet and Princeton WordNet. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2413–2418, Portorož, Slovenia. European Language Resources Association (ELRA).

Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. 2015. EVALution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 64–69, Beijing, China. Association for Computational Linguistics.

Erik Tjong Kim Sang. 2007. Extracting hypernym pairs from the web. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 165–168, Prague, Czech Republic. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. HyperLex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4):781–835.