清华大学

博士资格考试

题目：词义消歧任务分析和综述

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2023 年 6 月 23 日

目录

1 课题来源及研究的背景意义

1.1 课题的来源

1.2 课题研究背景及意义

一个词语，无论它是否出现在上下文中，都会有多种理解方式，即词汇的多义性。出于经济性原则，多义性现象在各种语言中普遍存在，并且早已受到语言学家的关注和重视[1-3]。词汇语义学指出，词汇的意义主要由指陈（reference）和内涵意义（sense）构成，前者表示该词语所指的事物，代表词类为具有具体谈论对象的指陈性名词、代词、动词等；后者则表示超出所指抽象的意义[4]，一般由较为抽象的名词（例如希望、热情），或是具有描述性质的形容词、副词组成。这两个种类构成了词汇多义性的两个方面。

多种因素都会导致词汇多义性，这往往和言语的历史发展相关。词汇语义在诞生之初一般都具有一个基本义，这往往是一种指陈义。之后言语社团的语言使用者通过联想、引申等手段，使同一形式的词汇片段产生了不同的意义。使用频繁的、用法固定的一些意义往往可以脱离上下文，成为该词汇的“永久性”意义，并作为单独一个义项为词典所吸纳，例如“手足”的身体部位义和兄弟义；另外一些依赖上下文甚至文化背景则被视作词汇的临时性意义，不同言语社团一般并未达成共识，也不作为单独义项罗列出来，例如“龙”在东方代表高贵，而在西方代表邪恶。

词汇多义性借由其所在的上下文往往可以得到消解，即得到一个明确的意义。计算语言学将这一任务定义为词义消歧（Word Sense Disambiguation），并试图通过机器学习的方法，训练出一个可以自动消歧的模型。常见的词义消歧模型将这一任务定义为多类分类任务（Multinominal classification），即从所有候选的词义选项中找出一个最佳的词义类别。分类任务默认词类标签之间是相互独立、语义正交的，这往往与真实的标签分布不一致；同时它也需要得知词汇的所有候选词义选项，而新兴的、未出现在语义词典中的词汇往往不具备这一条件。为了解决上述困难，词义消歧可以建模为一个结构化预测问题[1]（又被称为“定义建模”（Definition Modeling）），即让生成模型解码出代表意义的短语或者句子（一串词汇的序列）。尽管生成式建模更加贴近人类的直觉，但是由于复杂的模型结构，其需要更多的训练数据和训练开销。

词义消歧任务往往只针对实词（即名词、动词、形容词、副词）的“长久性”意义[2]，与词汇多义性相关的工程问题还包括：（1）共指消解 (Coreference Resolution)：确定文本中指向相同实体的名词和代词；（2）命名实体识别（Named Entity Recognition）：确定文本中所提的专有名词属于哪个实体；（3）隐喻检测（Metaphor Detection）：检测文本中涉及到隐喻用法的词汇；（4）词汇对译（colexification）：将一种语言的特定词汇翻译

[1] 结构化预测是指模型预测出一个向量型的结构体，例如一个相互依赖的序列，而并非预测出一个标量。

[2] 区别于依赖上下文才可以理解的临时义。

到另一门语言。其中，前三种涉及到词汇的指陈意义或者临时意义；最后一种体现了跨语言之间群体对于同一词汇的语义概念划分不一致。

确定词汇在上下文中的语义在计算语言学背景下仍存在很多挑战。以词义消歧任务为例，受到上下文范围限制或者言语者的主观意图，词汇本身就可能有多种理解（即“一对多”的任务），这体现了词义选择的不确定性，这种不确定性进而影响到更大单位——句子的语义理解。另一方面，候选词义并非严格相互正交，而是一个连续统。如何切分语义的粒度对于类别标签的构建和模型训练影响很大。同时，同一词汇的不同语义形式在语料库中自然出现的分布并非均匀，一般来说，罕见词义的训练语料远低于常见语义的，这样会导致模型倾向于常见语义。总体来说，语义相对于词汇是一个不可直接观测的隐变量，其形式具有连续性、模糊性和主观性。

1.3 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：

$$P(\mathcal{Z}_w, w, c) = P(\mathcal{Z}_w|w, c)P(w, c) = P(\mathcal{Z}_w|w)P(s). \tag{1}$$

其中，$s = \{w, c\}$。对于上下文依赖语义而言：

$$P(\mathcal{Z}_w, w, c) = P(\mathcal{Z}_w|w, c)P(w, c) = P(\mathcal{Z}_w|w, c)P(s). \tag{2}$$

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

任务名称	w_i	$\mathcal{Z}$ 类别	$\mathcal{Z}_w \perp\!\!\!\perp c\|w$	$\mathcal{E}$	$\mathcal{D}$
词义消歧（判别式）	实词	候选词义集合	✓	$\{c_i\}$	✓
词义消歧（生成式）	实词	词汇序列	✓	$\{c_i\}$	✗
共指消解	代词	专有名词	✗	$\{c_i\}$	✗
命名实体识别	专有名词	专有名词	✓	$\{c_i, \mathcal{K}\}$	✗
隐喻检测	实词	-	✗	$\{c_i, \mathcal{K}, \mathcal{H}\}$	✗
词汇对译	源语言实词	目标语言实词	✓	$\{c_i, \mathcal{H}\}$	✗

表 1 不同词汇多义性任务的对比

- 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：

$$P(z, w|\mathcal{E}) = P(z|w, \mathcal{E})P(w|\mathcal{E}). \tag{3}$$

1.3.1 词义消歧任务

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。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 $z = \max_i P(z_i|w, c_i)$。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，$\max P(g_1, ..., g_K|w, c_i)$。本文主要研究词义消歧任务。

1.3.2 其他任务

与词汇多义性相关的任务在目标词 w_i 的类型，词汇语义 $\mathcal{Z}$ 的类别，是否为上下文无关词义集，环境 $\mathcal{E}$ 类别，以及是否需要语义词典 $\mathcal{D}$ 上都与词义消歧任务有差异，表 1罗列了具体的差别。

2 国内外在词义消歧上的研究现状及分析

词义消歧任务尽管较为明确，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。

语料库	文章	语料句子	语料单词	标注	词义种类	词种类[3]	多义度
Senseval-2[5]	3	242	5,766	2,282	1,335	1,093	5.4
Senseval-3[6]	3	352	5,541	1,850	1,167	977	6.8
SemEval-07[7]	3	135	3,201	455	375	330	8.5
SemEval-13[8]	13	306	8,391	1,644	827	751	4.9
SemEval-15[9]	4	138	2,604	1,022	659	512	5.5
SemCor[10]	352	37,176	802,443	226,036	33,362	22,436	6.8
OMSTI[11]	-	813,798	30, 441, 386	911,134	3,730	1,149	8.9
WNGC[12]	-	-	1,621,000	449,000	-	-	-
OntoNotes[12]	-	-	1,500,000	-	-	-	-

表 2 不同语料库的数量统计对比

2.1 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 2展示了相关的数量统计对比。

2.1.1 SemCor

英文版的 SemCor[10] 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[13] 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[14]。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[15]。

2.1.2 OMSTI

OMSTI (One Million Sense-Tagged Instances)[11] 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[16], 利用外部软件（GIZA++[17]）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。

[3] 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。

2.1.3 WNGC

WNGC (WordNet Gloss Corpus)[4]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。

2.1.4 OntoNotes

OntoNotes 5.0[5]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。

2.1.5 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[18] 对它们进行了整理。它一般由五次比赛构成，分别为：（1）Senseval-2[5]；（2）Senseval-3[6] 的任务 1；（3）SemEval-07[7] 的任务 17；（4）SemEval-13[8] 的任务 12；（5）SemEval-15[9] 的任务 13。

2.2 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 3。

2.2.1 WordNet

WordNet[15] 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[6]。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

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

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

[6] 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.

知识库	词条数目	同义词集合数量	单义词数量	多义度[11]
WordNet 3.0[12]	155,287	117,659	101,863	2.50
HowNet[13]	237,974 [14]	35,202 [15]	-	-
SyntagNet	-	71,025	-	-
BabelNet[20]	-	22,130,060	-	-
BabelNet_EN	-	13,964,713	-	-

表 3 不同类型的知识库的数量统计对比

言的 WordNet[7]。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[8]和台湾大学语言所研发的中文词网[9]。

2.2.2 知网

知网（HowNet)[19] 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。[10]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。

2.2.3 BabelNet

BabelNet[20] 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。

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

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

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

[10] 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。

[11] 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。

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

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

[14] 包含中英文

[15] 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 模型方法

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[21]	SC	WN	✓	✗	✓
		GlossBERT[22]	SC	WN	✓	✗	✗
		EWISE[23]	SC	WN	✓	✗	✓
		EWISER[24]	SC+G	WN	✓	✓	✓
		MLWSD[25]	SC	WN	✗	✗	✓
		MLWSD*	SC	WN	✓	✓	✓
		RTWE[26]	SC	WN	✓	✗	✗
		RTWE*	SC+G	WN	✓	✓	✗
	语义检索任务	BEM[27]	SC	WN	✓	✗	✗
		Z-reweight[28]	SC	WN	✓	✗	✓
		SACE[29]	SC	WN	✓	✗	✓
		SACE*	SC+G	WN	✓	✗	✓
		ARES[30]	SC+WK	WN	✓	✗	
	截取式任务	ESCHER[31]	SC	WN		✗	✗
		ConSec[32]	SC	WN	✓	✗	✗
		ConSec*	SC+G	WN	✓	✓	✗
		KELESC[33]	SC	WN	✓	✓	✓
	生成式任务	Vec2Gloss[34]	-	-	✓	✗	✗
		Generationary[35]	CHA+SEM	-	✓	✗	✗
半/无监督	-	WSD_TM[36]	WK	-	✓	✓	✓
		WSD_LSA[37]	SE10	-	✗	✗	✗
完全知识驱动	基于相似性	Lesk[38]	-	WN	✓	✓	✗
		Lesk_ext[39]	-	WN	✓	✓	✓
		SREF[40]	-	WN	✓	✓	✓
	基于图算法	UKB[41]	-	WN+ESWN+EXWN	✓	✓	✓
		Babelfy[42]	-	BN	✗	✗	✓
		SyntagRank[43]	-	WN	✓	✓	✓
		WSDG[44]	-	WN+BN	✓	✗	✓

表 4 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。

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。

2.3.1 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 4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。

2.3.1.1 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：

$$score : \mathcal{Z} \times \mathcal{Z} \rightarrow [0, 1]. \tag{4}$$

不同的方法考虑不同的相似性方法。Rada 等人[45] 以及 Leacock 和 Chodorow[46] 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[38] 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[39] 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[] 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[40] 利用 BERT[47] 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。

2.3.1.2 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[41] 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[42] 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[43]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[48]。它采用的图算法仍然是个性化 PageRank 算法。WSDG[44] 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。

2.3.1.3　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[49]，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[50]，隐马尔可夫模型[51]，基于类别的概率[52] 和贝叶斯网络[53]。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。

2.3.2 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：

$$p_\theta(\mathcal{Z}|c, w, \mathcal{E}). \tag{5}$$

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：

$$1nn(q_\alpha(w, c, \mathcal{E}), k_\beta(\mathcal{Z})). \tag{6}$$

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。

2.3.2.1　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[54]，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[55]。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[10]，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[56]、朴素贝叶斯模型[57]、（浅层）神经网络[58]、k 近邻算法[59]、SVM 算法[60] 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2.2.1WordNet 的介绍）。GAS[21] 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[22] 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[23] 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[24] 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[25] 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[26] 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。

2.3.2.2 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[27] 利用两个编码器分别编码语义句子和上下文；之后有工作[28] 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[29] 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[30] 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。

2.3.2.3 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[31] 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[32] 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[33] 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。

2.3.2.4　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[61] 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[62]，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[34]，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[62]，长短时注意力机制[63]，到现在基于自注意力机制的 Transformer[64] 来实现。近期的做法[35] 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[63] 或者连续[65] 的隐变量生成模型去学习，或者用近似的分布[66] 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[67-69]。

2.3.3　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[70]，Glove[71]；上下文表示，例如 Bert[47]。规约的算法可以选择聚类算法[72]，或者隐变量模型[36-37]。

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[73] 和自训练[74] 两种方式。也有方法利用不确定性采样和主动学习[75] 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[76]，或者大语言生成模型[77]，例如 GPT-2 等方式。

2.4　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2.1.5）。为了统一标准和简化测试流程，研究者[18] 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[16]进行评测。表 5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：

[16] 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[78]。本文仅仅考虑 micro F1 评分。

2.4.1 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。

2.4.2 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。

2.4.3 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[17]。还有一个方法是 ChatGPT[18]，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。

表 5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 4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。

3 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[19]将两方面进行结合，做出了一些初步探讨。（论文参见附录 1）

[17] 根据 WordNet 的编纂特点，第一个义项是根据词频决定的。

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

[19] 现已被 Findings: Association Computational Linguistics 2023 会议录取。

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[19] 作为语义词典，它将语义分解为更小的单位：义素，很多方法[79-80] 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[81]，或者以现代汉语词典标注[82] 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[82-84] 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。

方法	SE02	SE03	SE07	SE13	SE15	ALL	名	动	形	副
ITA[24]	-	-	-	-	-	80.0				
LB_Mono[20]	-	-	-	-	-	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[21]	72.2	70.5	-	67.2	72.6	70.6	72.2	57.7	76.6	85.0
GlossBERT[22]	77.7	75.2	72.5	76.1	80.4	77.0	79.8	67.1	79.6	87.4
EWISE[23]	73.8	71.1	67.3	69.4	74.5	71.8	74.0	60.2	78.0	82.1
EWISER[24]	80.8	79.0	75.2	80.7	81.8	80.1	82.9	69.4	83.6	87.3
MLWSD[25]	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[26]	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[27]	79.4	77.4	74.5	79.7	81.7	79.0	81.4	68.5	83.0	87.9
Z-reweight[28]	79.6	76.5	71.9	78.9	82.5	78.6	-	-	-	-
SACE[29]	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[30]	78.0	77.1	71.0	77.3	83.2	77.9	80.6	68.3	80.5	83.5
ESCHER[31]	81.7	77.8	76.3	82.2	83.2	80.7	83.9	69.3	83.8	86.7
ConSec[32]	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[33]	82.2	78.1	76.7	82.2	83.0	81.2	84.3	69.4	84.0	86.7
Generationary[35]	77.8	73.7	68.8	78.3	77.6	76.3	79.8	63.3	80.1	84.7
Lesk_ext[39]	58.4	59.4	-	-	-	-	-	-	-	-
SREF[40]	72.7	71.5	61.5	76.4	79.5	73.5	78.5	56.6	79.0	76.9
UKB[41]	59.7	57.9	41.7	-	-	-	-	-	-	-
Babelfy[42]	-	68.3	62.7	65.9	-	-	-	-	-	-
SyntagRank[43]	71.6	72.0	59.3	72.2	75.8	71.7	64.1	-	-	-
WSDG[44]	68.7	68.3	58.9	66.4	70.7	67.7	71.1	51.9	75.4	80.9

表 5 各类消歧方法的性能对比

主要参考文献

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

[2] 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.

[3] 张志毅张庆云. 词汇语义学[M]. [出版地不详]: 北京: 商务印书馆, 2012.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

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

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

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

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

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

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

[51] 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.

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

[53] 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.

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

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

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

[57] 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.

[58] 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.

[59] 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.

[60] 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.

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

[62] 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.

[63] 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.

[64] 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.

[65] 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.

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

[67] 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.

[68] 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.

[69] 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.

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

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

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

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

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

[86] 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.

1 附录

论文题目[21]为：歧义性和不确定性的结合：词义消歧任务中的不确定性评估

论文摘要（中文）：词义消歧旨在根据上下文确定目标词的适当意义，这一任务对于自然语言理解至关重要。现有的监督方法将词义消歧视为分类任务，并取得了最优的性能。然而，它们忽略了现实世界中的词义选择的不确定性问题，这一问题往往是由于真实数据带有噪声并超出分布范围所导致的。本文广泛研究了针对词义消歧任务设计的基准测试中的不确定性估计。具体而言，针对一个算法先进的词义消歧模型，我们首先比较了四种不确定性分数，并验证了在模型的最后一层获得的传统预测概率不足以量化不确性。然后，我们通过选定的不确定性估计分数检验了模型捕捉数据和模型不确定性的能力，并发现模型充分反映了数据不确定性但低估了模型不确定性。此外，我们还探讨了许多词汇属性对数据不确定性的内在影响，并对以下四个关键方面进行了详细分析：句法类别、形态学、词义粒度和语义关系。

[21] 论文已经被 Findings: Association Computational Linguistics 2023 会议录取，预印版访问：https://arxiv.org/abs/2305.13119。

Ambiguity Meets Uncertainty: Investigating Uncertainty Estimation for Word Sense Disambiguation

Zhu Liu
Tsinghua University
School of Humanities
liuzhu22@mails.tsinghua.edu.cn

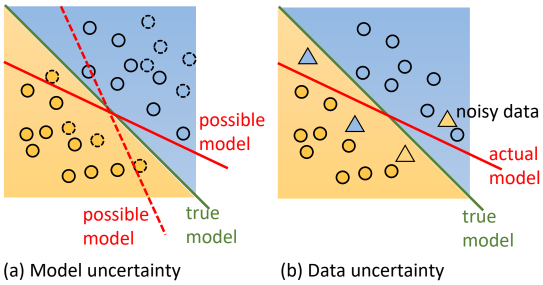
Ying Liu
Tsinghua University
School of Humanities
yingliu@tsinghua.edu.cn

Abstract

Word sense disambiguation (WSD), which aims to determine an appropriate sense for a target word given its context, is crucial for natural language understanding. Existing supervised methods treat WSD as a classification task and have achieved remarkable performance. However, they ignore uncertainty estimation (UE) in the real-world setting, where the data is always noisy and out of distribution. This paper extensively studies UE on the benchmark designed for WSD. Specifically, we first compare four uncertainty scores for a state-of-the-art WSD model and verify that the conventional predictive probabilities obtained at the final layer of the model are inadequate to quantify uncertainty. Then, we examine the capability of capturing data and model uncertainties by the model with the selected UE score on well-designed test scenarios and discover that the model adequately reflects data uncertainty but underestimates model uncertainty. Furthermore, we explore numerous lexical properties that intrinsically affect data uncertainty and provide a detailed analysis of four critical aspects: the syntactic category, morphology, sense granularity, and semantic relations. The code is available at https://github.com/RyanLiut/WSD-UE.

1 Introduction

Disambiguating a word in a given context is fundamental to natural language understanding (NLU) tasks, such as machine translation (Gonzales et al., 2017), question answering (Ferrández et al., 2006), and coreference resolution (Hu and Liu, 2011). This task of word sense disambiguation (WSD) targets polysemous or homonymous words and determines the most appropriate sense based on their surrounding contexts. For example, the ambiguous word *book* refers to two completely distinct meanings in the following sentences: i)“*Book* a hotel, please.”, ii) “Read the *book*, please”. The phenomenon is universal to all languages and has been paid much attention since the very beginning of artificial intelligence (AI) (Weaver, 1952).



(a) Model uncertainty

(b) Data uncertainty

Figure 1: Two types of uncertainties in the case of classification. The green line indicates the true model (decision boundary), while the red shows possible models. Circles and triangles with different colors illustrate clean and noisy data with corresponding labels.

Existing supervised methods (Blevins and Zettlemoyer, 2020; Conia and Navigli, 2021; Bevilacqua and Navigli, 2020; Calabrese et al., 2021; Huang et al., 2019) cast WSD as a classification task in which a neural networks (NNs)-based classifier is trained from WordNet (Miller et al., 1990), a dictionary-like inventory. Although they have achieved the state of the art on WSD benchmarks, with some even breaking through the estimated upper bound on human inter-annotator agreement in terms of accuracy (Bevilacqua and Navigli, 2020), they do not capture or measure uncertainty. Uncertainty estimation (UE) answers a question as follows: *To what extent is the model certain that its choices are correct?* A model can be unsure due to the noisy or out-of-domain data, especially in a real-world setting. This estimation delivers valuable insights to the WSD practitioners since we could pass the input with high uncertainty to a human for classification.

UE is an essential requirement for WSD. Interestingly, the word “ambiguous” (in terms of the task of word sense *disambiguation*) itself is ambiguous: it refers to i) doubtful or uncertain especially from obscurity or indistinctness, and ii) capable of being understood in two or more possible senses or ways, according to the Merriam-Webster dictionary[1]. The conventional treatment only considers its second aspect but disregards the first uncertainty-related sense. In reality, there are many situations where uncertainties arise (Yarin, 2016). The first situation assumes a true model to which each trained model approximates. Uncertainty appears when the structures and parameters of the possible models vary; we refer to it as model uncertainty (Figure 1 (a)) in this paper. *Model uncertainty* can be reduced when collecting enough data, i.e., adequate knowledge to recognize the true model and out-of-distribution (OOD) data is always used to test model uncertainty. It has been observed that WSD is prone to domain shift and bias towards the most frequent sense (MFS) (Raganato et al., 2017). Therefore, it is essential to quantify model uncertainty in the task.

Another uncertainty is related to the data itself and cannot be explained away, which is referred to as *data uncertainty* (also called aleatoric uncertainty). Data uncertainty happens when the observation is imperfect, noisy, or obscure (Figure 1 (b)). Even if there is enough data, we cannot obtain results with high confidence. WSD is context-sensitive, and the model output could be divergent due to partial or missing context. Even worse, some words have literal and non-literal meanings and can be understood differently. With a fine-grained WordNet (Miller et al., 1990) as a reference inventory, the inter-annotator disagreement is up to 20% to 30% (Navigli, 2009): even human annotators cannot agree on the correct sense of these words.

In this paper, we perform extensive experiments to assess the uncertainty of a SOTA model (Conia and Navigli, 2021) on WSD benchmarks. First, we compare the probability of the model output with the other three uncertainty scores and conclude that this probability is inadequate to UE, which is consistent with previous research (Gal and Ghahramani, 2016). Then, with the selected score, we evaluate data uncertainty in two designed scenarios: window-controlled and syntax-controlled contexts, which simulate noisy real-world data. Further, we estimate model uncertainty on an existing OOD dataset (Maru et al., 2022) and find that the model underestimates model uncertainty compared to the adequate measure of data uncertainty. Finally, we design an extensive controlled procedure to determine which lexical properties affect uncertainty estimation. The results demonstrate that morphology (parts of speech and number of morphemes), inventory organization (number of annotated ground-truth senses and polysemy degree) and semantic relations (hyponym) influence the uncertainty scores.

2 Related Work

2.1 Word Sense Disambiguation

Methods of WSD are usually split into two categories, which are knowledge-based and supervised models. Knowledge-based methods employ graph algorithms, e.g., clique approximation (Moro et al., 2014), random walks (Agirre et al., 2014), or game theory (Tripodi and Navigli, 2019) on semantic networks, such as WordNet (Miller et al., 1990), BabelNet (Navigli and Ponzetto, 2012). These methods do not acquire much annotation effort but usually perform worse than their supervised counterpart due to their independence from the annotated data. Supervised disambiguation is data-driven and utilizes manually sense-annotated data sets. Regarding each candidate sense as a class, these models treat WSD as the task of multi-class classification and utilize deep learning techniques, e.g., transformers (Conia and Navigli, 2021; Bevilacqua and Navigli, 2019). Some also integrate various parts of the knowledge base, such as neighboring embeddings (Loureiro and Jorge, 2019), relations (Conia and Navigli, 2021), and graph structure (Bevilacqua and Navigli, 2020). These methods have achieved SOTA performance and even broken through the ceiling human could reach (Bevilacqua and Navigli, 2020). However, these methods treat disambiguation as a deterministic process and neglect the aspect of uncertainty.

2.2 Uncertainty Estimation

Uncertainty estimation (UE) has been studied extensively, especially in computer vision (Gal et al., 2017) and robust AI (Stutz, 2022). Methods capture uncertainty in a Bayesian or non-Bayesian manner. Bayesian neural networks (Neal, 2012) offer a mathematical grounded framework to model predictive uncertainty but usually comes with prohibitive inference cost. Recent work proved MC Dropout approximates Bayesian inference in deep Gaussian Processes and has been widely applied

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

in many UE applications (Vazhentsev et al., 2022; Kochkina and Liakata, 2020) due to its simplicity. During recent years, the field of natural language processing has witnessed the development of an increasing number of uncertain-aware applications, such as Machine Translation (Glushkova et al., 2021), Summarization (Gidiotis and Tsoumakas, 2021) and Information Retrieval (Penha and Hauff, 2021). Nevertheless, little attention has been paid to the combination of UE and WSD. An early work (Zhu et al., 2008) explored uncertainty to select informative data in their active learning framework. However, the uncertainty estimation for WSD is not explored extensively, as we do in a quantitative and qualitative way.

3 Uncertainty Scenarios

3.1 Problem Formulation

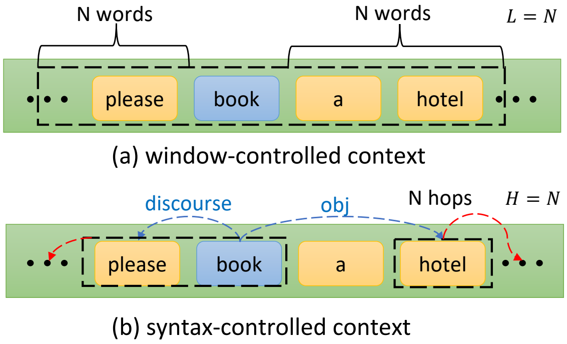
Given a target word w_i in a context $c_i = (w_0, w_1, ..., w_i, ..., w_W)$ of W words, a WSD model selects the best label $\hat{y}_i$ from a candidate sense set $S_i = (y_1, y_2, ..., y_M)$ consisting of M classes. A neural network p_θ with the parameter θ usually obtains a probability p_i over M classes by a softmax function which normalizes the model output f_i:

$$p_i = \text{SoftMax}(f_i(w_i|c_i;\theta)). \tag{1}$$

During training, the probability is used to calculate cross-entropy loss, which can be recognized as a probability for each candidate class during the inference. Such a point estimation of model function has been erroneously interpreted as model confidence (Gal and Ghahramani, 2016). The goal of UE is to find a suitable p_i to better reflect true predictive distribution under data and model uncertainty sources. Suppose we have a reasonable score $s(p_i) \in \mathcal{S}$ indicating UE, where $\mathcal{S}$ is a metric space, we expect $s^a > s^b$ when a situation a is more uncertain than b.

3.2 Data Uncertainty: Controllable Context

Data uncertainty measures the uncertainty caused by imperfect or noisy data. We consider that such noises could happen in the context surrounding the target word, considering WSD is a context-sensitive task. With different degrees of missing parts in the context, the model is expected to obtain predictions with different qualifications of uncertainty. To simulate this scenario, we control the range of context based on two signals: the window and the syntax, as illustrated in Figure 2.



(a) window-controlled context

(b) syntax-controlled context

Figure 2: Two types of controlled context in the data uncertainty setting. The target word is highlighted in blue. The box with a black dotted line shows the final chosen context. We show the dependency relation in blue and red.

3.2.1 Window-controlled Context

We choose L words both on the left and right of the target word w_i as the window-controlled context $c_L^{\text{WC}} = (w_l, w_{i-1}, w_i, w_{i+1}, ..., w_h)$, where $l = \max(i - L, 0)$ and $h = \min(i + L, W)$ are the lower index and the higher index. With a hypothesis that longer context tends to contain more clues to disambiguate a word and a suitable UE score s, we expect that $s_a^{\text{WC}} > s_b^{\text{WC}}$, where two window-controlled contexts are extracted with the length of a and b, and $a < b$.

3.2.2 Syntax-controlled Context

In our second controlled method, we utilize the neighboring syntax around w_i. Specifically, we parse the universal syntactical dependency relations between words using tools of Stanza (Qi et al., 2020). This is represented as a form of graph structure $\mathcal{G} = (\mathcal{N}, \mathcal{R})$, where $\mathcal{N}$ denotes the nodes, i.e., each word, and $\mathcal{R} =< n^h, n^t, r >$ is the relation r from the head node n^h to tail node n^t. For example, when r is *nsubj*, that means n^h is the subject of n^t. We iteratively obtain a syntax-related [2] neighboring set with the H hops of the target word w_i as c_H^{DP} in the following approach. Initially, c_H^{DP} only contains w_i. After one hop, c_H^{DP} collects the head node and tail nodes of w_i. The procedure is repeated H times, with more syntactically related words added. We also rationally hypothesize a smaller s^{DP}, which measures uncertainty under

[2] We denote this scenario as DP, since we utilize dependency parsing as the syntactic representation.

syntax-controlled context, favors the context with a larger H. We highlight that the syntax-controlled context leverages the nonlinear dependency distance (Heringer et al., 1980) between words in connection, compared to the linear distance in the scenario of window-controlled context.

3.3 Model Uncertainty: OOD Test

Model uncertainty is another crucial aspect of UE, widely studied in the machine learning community. Lacking knowledge, models with different architectures and parameters could output indeterminate results. Testing a model on OOD datasets is a usual method to estimate model uncertainty. In the task of WSD, we employ an existing dataset 42D (Maru et al., 2022) designed for a more challenging benchmark. This dataset built on the British National Corpus is challenging because 1) for each instance, the ground truth does not occur in SemCor (Miller et al., 1994), which is the standard training data for WSD, and 2) is not the first sense in WordNet to avoid most frequent sense bias issue (Campolungo et al., 2022). 42D also has different text domains from the training corpus. These confirm that 42D is an ideal OOD dataset.

4 Experiments

4.1 Model and Datasets

We conduct our UE for a SOTA model MLS (Conia and Navigli, 2021), with the best parameters released by the authors. They framed WSD as a multi-label problem and trained a BERT-large-cased model (Kenton and Toutanova, 2019) on the standard WSD training dataset SemCor (Miller et al., 1994). We follow their settings except for using Dropout during inference when performing Monte Carlo Dropout (MC Dropout). We set the number of samples T to be 20, conduct 3 rounds, and report the averaged performance.

As regards the evaluation benchmark, we use the Unified Evaluation Framework for English all-words WSD proposed by (Raganato et al., 2017). This includes five standard datasets, namely, Senseval-2, Senseval-3, SemEval-2007, SemEval-2013, and SemEval-2015. The whole datasets concatenating all these data with different parts of speech (POS) are also evaluated. Note that in our second part, We use a portion of SemEval-2007 to investigate data uncertainty and 42D is used for model uncertainty.

4.2 Uncertainty Estimation Scores

We apply four methods as our uncertainty estimation (UE) scores. One trivial baseline (Geifman and El-Yaniv, 2017) regards the Softmax output p_i as the confidence values over classes $y = s \in S$. We calculate the uncertainty score based on the maximum probability as $u_{\text{MP}}(x) = 1 - \max_{s \in S} p(y = s|x)$.

The other three methods are based on MC Dropout, which has been proved theoretically as approximate Bayesian inference in deep Gaussian processes (Gal and Ghahramani, 2016). Specifically, we conduct T stochastic forward passes during inference with Dropout random masks and obtain T probabilities p_t. Following the work (Vazhentsev et al., 2022), we use the following measures:

- Sampled maximum probability (SMP) takes the sample mean as the final confidence before an MP is applied: $u_{\text{SMP}} = 1 - \max_{s \in S} \frac{1}{T} \sum_{t=1}^{T} p_t^s$, where p_t^s refers to the probability of belonging to class s at the $t'th$ forward pass.

- Probability variance (PV) (Gal et al., 2017) calculates the variance before averaging over all the class probabilities: $u_{\text{PV}} = \frac{1}{S} \sum_{s=1}^{S} \left(\frac{1}{T} \sum_{t=1}^{T} (p_t^s - \overline{p^s})^2 \right)$.

- Bayesian active learning by disagreement (BALD) (Houlsby et al., 2011) measures the mutual information between model parameters and predictive distribution: $u_{\text{BALD}} = - \sum_{s=1}^{S} \overline{p^s} \log \overline{p^s} + \frac{1}{T} \sum_{s,t} p_t^s \log p_t^s$.

Note that these scores are instance-specific and we report the averaged results over all the samples.

4.3 Metrics on UE scores

While UE scores are a measure of uncertainty, we also need metrics to judge and compare the quality of different UE scores. A hypothesis is that a sample with a high uncertainty score is more likely to be erroneous and removing such instances could boost the performance. We employ two metrics following the work (Vazhentsev et al., 2022): area under the risk courage curve (**RCC**) (El-Yaniv et al., 2010) and reversed pair proportion (**RPP**) (Xin et al., 2021). RCC calculates the cumulative sum of loss due to misclassification according to the uncertainty level for rejections of the predictions.

UE Score	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

Table 1: UE score comparisons on five standard WSD datasets.

UE Score	NOUN		VERB		ADJ		ADV		ALL	
	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

Table 2: UE score comparisons on all the datasets with different kinds of POS.

A larger RCC indicates that uncertainty estimation negatively impacts the classification. Note that we use the normalized RCC by dividing the size of the dataset. RPP counts the proportion of instances whose uncertainty level is inconsistent with its loss level compared to another sample. For any pair of instances x_i and x_j with their UE score $u(x)$ and loss value $l(x)$:

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

where n is the size of the dataset.

5 Results and Analysis

In the first part, we show the quantitative results of different UE scores and the performances of data and model uncertainty. Then a qualitative result demonstrates specific instances with a range of uncertainties. This motivates us to analyze which lexical properties mainly affect uncertainty in the last part.

5.1 Quantitative Results

5.1.1 Which UE score is better?

We measure the four UE scores, MP, SMP, PV, and BALD in terms of two metrics, RCC and RPP. The results of five standard datasets are shown in Table 1 while the performance on all the datasets involving different parts of speech is demonstrated in Table 2. For most of the data, SMP outperforms the other three scores in spite of some inconsistent results where MP has a slight advantage, such as on SemEval-15. Interestingly enough, softmax-based scores i.e., MP and SMP, surpass the other two, PV and BALD. Similar results can be observed in the work (Vazhentsev et al., 2022). This may be due to the fact that the former scores are directly used as the input of the maximum likelihood objective, thus more accurately approximating the real distribution.

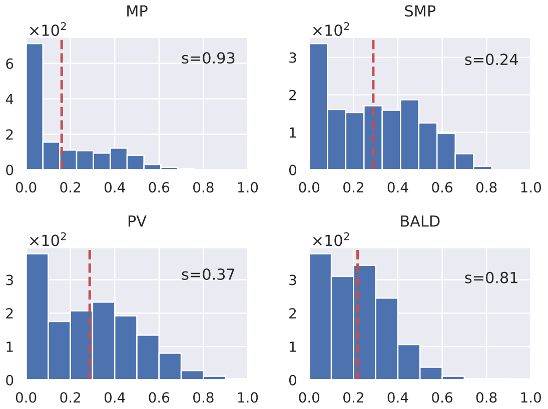


Figure 3: The distribution of four UE scores on misclassified instances of all datasets. A red dotted line indicates the average value. We calculate the sample skewness s for each score as well. Note that PV and BALD scores are normalized into the range from 0 to 1.

To further investigate the distribution of these four scores, we show the histograms of these scores in the misclassified instances, as illustrated in Figure 3. We also display the averaged value (a red dotted line) and the sample skewness s, calculated as the Fisher-Pearson coefficient (Zwillinger and Kokoska, 1999). Since here we focus on the misclassified samples, the cases of all the samples and those correctly classified are reported in Appendix A.1. This shows that MP has a more long-tailed and skewed distribution than scores based on MC Dropout, indicating MP is overconfident towards the wrong cases. However, the other three metrics have a more balanced distribution. This verifies the common concern on the SoftMax output of a single forward as an indication of confidence.

Finally, given its outstanding performance, we chose SMP as our uncertainty score in the following experiments.

5.1.2 How does the model capture data uncertainty?

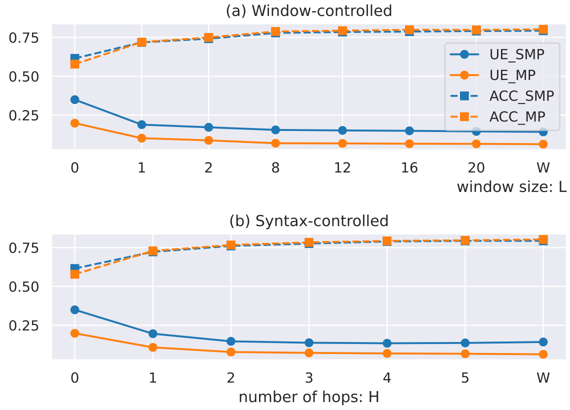


Figure 4: UE scores (SMP and MP) and accuracy (F1 score) vary depending on the range of context for (a) window-controlled setting and (b) syntax-controlled setting. Note that "0" indicates that only target words without context are available to the model. On the other hand, "W" means the whole context is available.

We verify data uncertainty in window-controlled and syntax-controlled scenarios, as shown in Figure 4. In the first setting, UE becomes less, and the accuracy grows with the increase of window size T. This indicates that the model perceives more and more confidence in the data, accessible to more neighboring words. The trend is similar in the syntax-controlled setting. These show that the model can adequately capture data uncertainty. SMP has a larger uncertainty than MP, especially in a sparse context, such as L or H is equal to 0 or 1, where the model is expected to be much more uncertain. We report the comparison of the other two sample-based scores, PV and BALD in Appendix A.2.

5.1.3 How does the model capture model uncertainty?

We examine the model uncertainty on the 42D dataset in Figure 5. The result shows OOD dataset is indeed a challenging benchmark for WSD. However, even with worse performance, the model fails to give a high UE score. We compare it with the most uncertain cases but similar accuracy in the settings of data uncertainty, i.e., without any context when $L = 0$. The OOD setting has a lower level of uncertainty, especially in the misclassified samples, even if it has degraded performance. This implies that the model underestimates the uncertainty level in model uncertainty. We show the performance of MP, PV, and BALD in Appendix A.3.

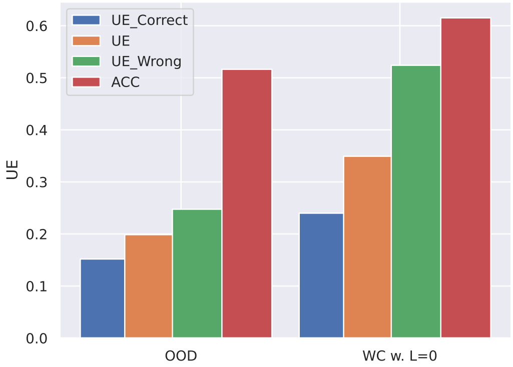


Figure 5: Uncertainty and accuracy (F1) scores for model uncertainty (OOD) and data uncertainty (controlled context) scenarios. We use window-controlled UE with L=0 (WC w. L=0). It is evaluated in all the data instances and wrongly (UE_Wrong) or correctly (UE_Correct) classified instances.

5.2 Qualitative Results

To investigate what kinds of words given a context tend to be uncertain, we obtain the final UE score for each word by averaging SMP scores for instances sharing the same form of lemma. In Figure 6, We show the word clouds for words with the most uncertain (left (a)) and certain (right (b)) meanings. We remove some unrepresented words whose number of candidate senses is less than 3. With respect to the most uncertain lemmas, there are words such as *settle*, *cover* etc. Most of them are verbs and own multiple candidate senses. As for most certain cases, the senses of nouns like *bird*, *bed*, and *article* are determined with low uncertainty. These phenomena motivate us to investigate which lexical properties affect uncertainty estimation in the next part. It is noted that we concentrate on data uncertainty instead of model uncertainty, based on the investigation in Subsection 5.1, which appears due to the data itself, i.e., lexical characteristics.

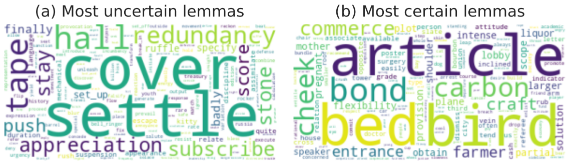


Figure 6: Word clouds for lemmas where a larger font indicates higher (a) or lower (b) UE scores.

5.3 Effects on Uncertainty

We explore which lexical properties affect uncertainty estimation from four aspects: the syntactic category (Folk and Morris, 2003), morphology [3](Lieber, 2004), sense granularity and semantic relations (Sternefeld and Zimmermann, 2013), motivated by linguistic and cognitive studies. Regarding syntactic categories, we focus on four i.e., parts of speech (**POS**) for target content words. Morphology aims at the number of morphemes (**nMorph**). A sense inventory refers to the sense items in a dictionary, whose granularity influences the candidate sense listing for the target word and its sense annotation (Kilgarriff, 1997). We consider two aspects:

- number of annotated ground-truth senses (**nGT**);
- number of candidate senses, i.e., polysemy degree (**nPD**);

To consider semantic interactions with other words, we utilize WordNet (Miller et al., 1990), a semantic network to extract lexical relations. Specifically, we concentrate on the hyponym and synonymy relations. A word (or sense) is a hyponym of another if the first is more specific, denoting a subclass of the other. For example, *table* is a hyponym of *furniture*. Each word as a node in WordNet lies in a hyponym tree, where the depth implies the degree of specification, denoted as **dHypo**. Meanwhile, we also explore the size of the synonymy set (**dSyno**) into which the ground-truth sense falls.

We perform linear regression analysis and conclude that most effects are significant as coefficients to the UE score, except for dSyno and ADV of POS. This is consistent with our result in Subection 5.3.3. The summary of the linear regression is shown in Appendix A.4. Afterwards, we design a controlled procedure to analyze and balance different effects. First, samples are drawn from all the test instances depending on some conditions, including nGT and POS. Afterward, we aggregate test data in one of three manners: *instance* (I), *lemma* (L), and *sense* (S) and average the UE values for the instances with the same manner. I represents each occurrence of the target word, L considers words with different inflections (e.g., *works* and *worked*), and S targets words with the same ground-truth sense. The sampled data is then grouped into N levels in terms of the values for the different effects in question. Finally, we calculate the mean UE score for each group and their corresponding T-test and p values. We heuristically set different choices of N for different effects, considering the trade-off of level granularity and sample sparsity. The p-value is expected to be lower than 5%. The overall comparison is summarized in Table 3 with the number and value range of different levels in Table 4.

5.3.1 Syntactic Category and Morphology

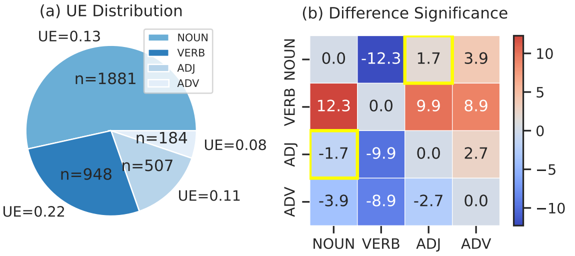


Figure 7: Averaged UE scores and numbers for instances aggregated by *sense*, with different parts of speech (a) and the corresponding difference significance for each pair (b). The heatmap (b) shows the T-test values where a higher absolute value (grids with a deeper color) indicates a more significant difference. We highlight the grid with a corresponding p value larger than 5%, implying no significant difference.

We show the averaged UE scores for instances with different POS and their corresponding T-test value in Figure 7. Except for the NOUN-ADJ pair, verbal instances are more significantly uncertain than NOUN or ADJ, while ADV has the least uncertainty. The result implies the senses of verbs are generally harder to determine than other categories, consistent with previous work (Barba et al., 2021; Campolungo et al., 2022). This is reflected in Table 2 and Figure 6.

[3] Here, we mainly consider derivational morphology. Multiword expressions e.g., compound words are included as well. Words with different inflectional morphology are regarded as the same lemma form.

Effect	Condition	Agg.	Uncertainty Estimation			Difference Significance		
			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

Table 3: Different uncertainty estimations (SMP) for different levels and corresponding difference significance (p values) of various effects involving morphology, inventory organization and semantic relations. Agg. means aggregation manners of the lemma (L), instance (I), and sense (S).

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

Table 4: The number and range of effects quantified into different levels for various effects.

We further explore the effects of morphology in Table 3. After extracting morphemes for each word using an off-line tool [4], we count the number of morphemes (denoted as nMorph). Since words with different parts of speech may have distinct mechanisms of word formation rules, we split data according to POS before averaging their UE scores and calculating corresponding difference significance. It shows that generally, the more morphemes a word consists of, the more uncertain its semantics would be. This is expected from the perspective of derivational morphology since adding prefixes, or suffixes could specify the stem words and have a relatively predictable meaning. For example, "V-ation" indicates the action or process of the stem verb, e.g., education, memorization. According to T-test in Table 3, UE scores of different levels for nouns are significantly distinct, while the difference is not so significant for other categories. It is because the derivational nouns including compound words are more representative and productive than other categories. This can be demonstrated by the fact that nouns contain the highest number of morphemes as shown in Table 4.

5.3.2 Sense Granularity

We first consider the number of ground-truth senses, i.e., nGT. During the annotation process, a not insignificant 5% of the target words is labeled multiple senses (Conia and Navigli, 2021). This reflects the difficulty in choosing the most appropriate meaning, even for human annotators. Given their contexts, the semantics of these words are expected to be more uncertain, and our result is consistent with this fact. We control nGT to be 1 in the remaining evaluation to eliminate its influence.

Second, we study the effect of polysemy degree (the number of possible candidates), i.e., nPD. It shows that target words with a more significant polysemy degree tend to be more uncertain. It is intuitively understandable because words with more possible meanings are always commonplace and easily prone to semantic change, e.g., *go*, *play*. Furthermore, their sense descriptions in WordNet are more fine-grained, indistinguishable in some cases even for humans. However, words with less polysemy degrees, such as compound words, are more certain in various contexts.

[4] https://polyglot.readthedocs.io/en/latest

5.3.3 Semantic relation

We discuss the effects of semantic relations for the target word in terms of WordNet. We first consider the hyponym relations, i.e., the depth in which a word node lies in the hyponym relation tree, as denoted by dHypo. Since nouns have clearer instances of hyponymy relation, we only consider this category. The results displayed in Table 3 show that instances with a deeper hyponym tend to own a certain meaning and the difference between each pair of levels is significant. That indicates that more specific concepts have a more determinate disambiguation, which is intuitive.

Another semantic relation is synonymy, as represented by dSyno. The measurement reveals that instances among different levels of the number of synonyms do not differ from each other significantly. This implies that whether the ground-truth meaning has more neighbors with similar semantics has less impact on the decision of uncertainty.

6 Conclusion

We explore the uncertainty estimation for WSD. First, we compare various uncertainty scores. Then we choose SMP as the uncertainty indicator and examine to what extent a SOTA model captures data uncertainty and model uncertainty. Experiments demonstrate that the model estimates data uncertainty adequately but underestimates model uncertainty. We further explore effects that influence uncertainty estimation in the perspectives of morphology, inventory organization and semantic relations. We will integrate WSD with uncertainty estimation into downstream applications in the future.

7 Limitations

Despite being easily adapted to current deep learning architectures, one concern about multiple-forward sampling methods is efficiency, since it has to repeat T processes to evaluate uncertainty in the stage of inference. We leave efficient variants of sampling methods for future work.

Another glaring issue is the focus on only English. Different languages may have different effects on uncertainty estimation due to e.g., distinct forms of morphology. Thus, some conclusions may vary according to the language in question. We hope that follow-up works will refine and complement our insights on a more representative sample of natural languages.

8 Ethics Statement

We do not foresee any immediate negative ethical consequences of our research.

9 Broader Impact Statement

Knowing what we do not know, i.e., a well-calibrated uncertainty estimation, is fundamental for an AI-assisted application in the real world. In the area of word sense disambiguation, the ambiguity and vagueness inherent in lexical semantics require a model to represent and measure uncertainty effectively. Our work explores the combination of these two areas and hopes that it will provide an approach to understanding the characteristics of languages.

10 Acknowledgements

The authors thank the anonymous reviewers for their valuable comments and constructive feedback on the manuscript. We also thank Rui Fang for his discussions on the linear regression analysis. This work is supported by the 2018 National Major Program of Philosophy and Social Science Fund “Analyses and Researches of Classic Texts of Classical Literature Based on Big Data Technology” (18ZDA238) and Tsinghua University Initiative Scientific Research Program (2019THZWJC38).

References

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021. Consec: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503.

Michele Bevilacqua and Roberto Navigli. 2019. Quasi bidirectional encoder representations from transformers for word sense disambiguation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 122–131.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017.

Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. 2021. Evilbert: Learning task-agnostic multimodal sense embeddings. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 481–487.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. Dibimt: A novel benchmark for measuring word sense disambiguation biases in machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352.

Simone Conia and Roberto Navigli. 2021. Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3269–3275.

Ran El-Yaniv et al. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5).

S Ferrández, Sandra Roger, Antonio Ferrández, Antonia Aguilar, and Pilar López-Moreno. 2006. A new proposal of word sense disambiguation for nouns on a question answering system. *Advances in Natural Language Processing. Research in Computing Science*, 18:83–92.

Jocelyn R Folk and Robin K Morris. 2003. Effects of syntactic category assignment on lexical ambiguity resolution in reading: An eye movement analysis. *Memory & Cognition*, 31:87–99.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

Alexios Gidiotis and Grigorios Tsoumakas. 2021. Uncertainty-aware abstractive summarization. *arXiv preprint arXiv:2105.10155*.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André FT Martins. 2021. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19.

Hans Jürgen Heringer, Bruno Strecker, and Rainer Wimmer. 1980. *Syntax: Fragen, Lösungen, Alternativen*. Fink.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *stat*, 1050:24.

Shangfeng Hu and Chengfei Liu. 2011. Incorporating coreference resolution into word sense disambiguation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 265–276. Springer.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuan-Jing Huang. 2019. Glossbert: Bert for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

Elena Kochkina and Maria Liakata. 2020. Estimating predictive uncertainty for rumour verification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6964–6981.

Rochelle Lieber. 2004. *Morphology and lexical semantics*, volume 104. Cambridge University Press.

Daniel Loureiro and Alipio Jorge. 2019. Language modelling makes sense: Propagating representations through wordnet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691.

Marco Maru, Simone Conia, Michele Bevilacqua, and Roberto Navigli. 2022. Nibbling at the hard core of word sense disambiguation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4724–4737.

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

George A Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G Thomas. 1994. Using a semantic concordance for sense identification. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

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

Radford M Neal. 2012. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Gustavo Penha and Claudia Hauff. 2021. On the calibration and uncertainty of neural learning to rank models for conversational search. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 160–170.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110.

Wolfgang Sternefeld and Thomas Ede Zimmermann. 2013. *Introduction to Semantics: An Essential Guide to the Composition of Meaning (Mouton Textbook)*. De Gruyter Mouton.

David Stutz. 2022. Understanding and improving robustness and uncertainty estimation in deep learning. *Saarländische Universitäts-und Landesbibliothek*.

Rocco Tripodi and Roberto Navigli. 2019. Game theory meets embeddings: a unified framework for word sense disambiguation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 88–99.

Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gusev, Mikhail Burtsev, et al. 2022. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8237–8252.

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

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. In *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)*, pages 1040–1051.

Gal Yarin. 2016. Uncertainty in deep learning. *University of Cambridge, Cambridge*.

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1137–1144.

Daniel Zwillinger and Stephen Kokoska. 1999. *CRC standard probability and statistics tables and formulae*. Crc Press.

A Appendix

A.1 Distribution of UE scores

We illustrate the distribution of UE scores, i.e., MP, SMP, PV and BALD for all the test samples in Figure 8 and samples that are correctly predicted in Figure 9. We assume samples that the model could accurately predict are easy and thus have a more certain meaning. Although SMP is not so long-tailed as MP in the case of correctly predicted samples, we do not expect a metric "overconfident" in all the cases, especially in the misclassified instances.

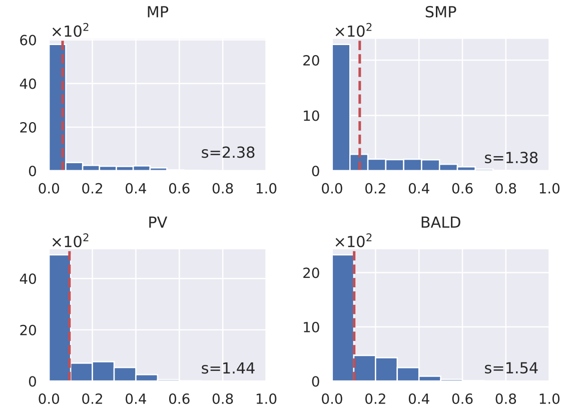


Figure 8: The distribution of four UE scores on all the test samples. The averaged value is indicated by a red dotted line. We calculate the sample skewness for each score as well.

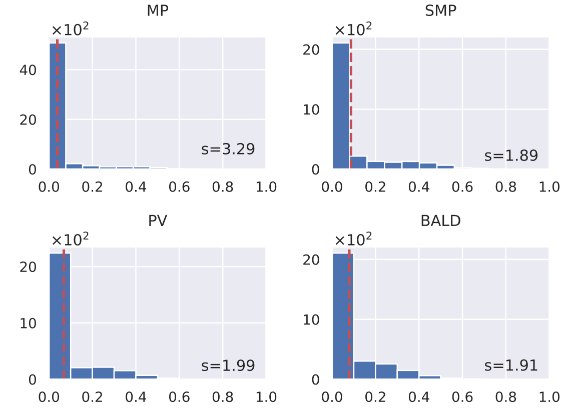


Figure 9: UE distribution on well-classified samples.

A.2 Other Scores for Data Uncertainty

We display the other two sample-based scores PV and BALD, in comparison with SMP in two data uncertainty scenarios in Figure 10. SMP has a higher uncertain score than the other two, especially in the more sparse context (e.g., $L = 0$), as we expected.

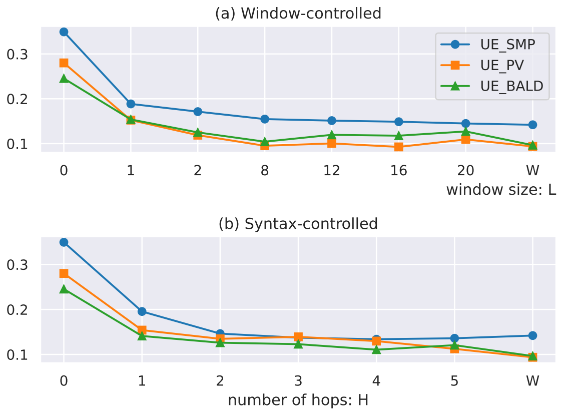


Figure 10: UE scores (SMP, PV, and BALD) vary depending on the range of context for (a) window-controlled setting and (b) syntax-controlled setting.

A.3 Other Scores for Model Uncertainty

We illustrate the other three UE scores (MP, PV and BALD) and accuracy for the scenario of model uncertainty compared with the least uncertain case for data uncertainty (L=0) in Figure 11, Figure 12 and Figure 13, respectively. The conclusion that UE scores underestimate model uncertainty is similar to that of MP.

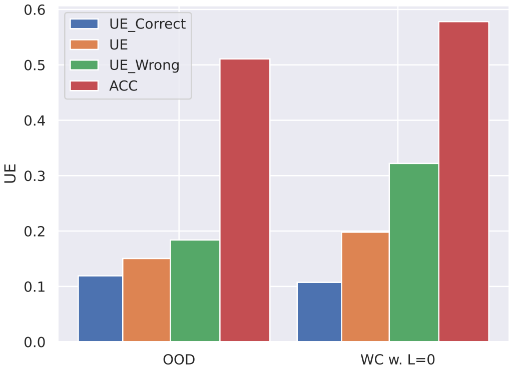


Figure 11: Uncertainty (MP) and accuracy scores for model uncertainty (OOD) and data uncertainty (controlled context) scenarios. We use window-controlled UE with L=0 (WC w. L=0). It is evaluated in all the data instances and wrongly (UE_Wrong) or correctly (UE_Correct) classified instances.

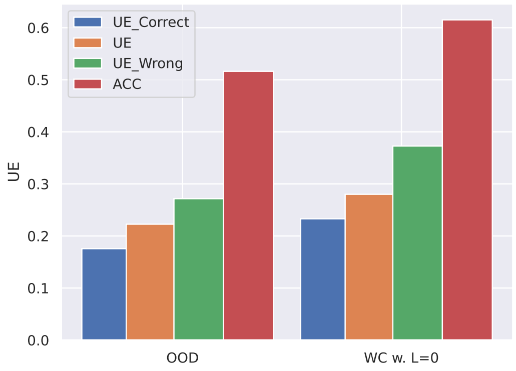


Figure 12: Uncertainty (PV) and accuracy scores for model uncertainty (OOD) and data uncertainty (controlled context) scenarios.

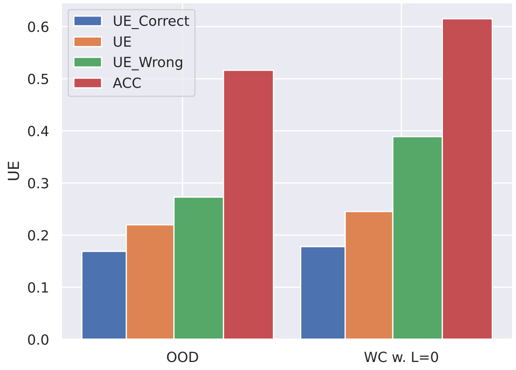
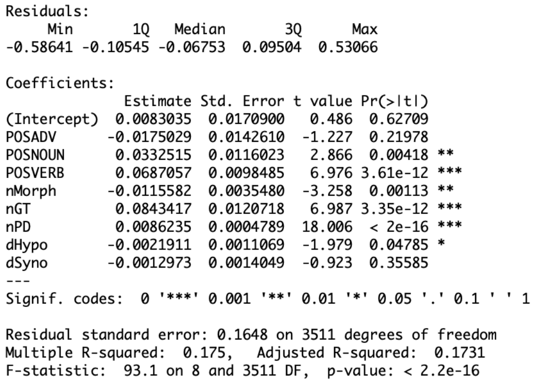


Figure 13: Uncertainty (BALD) and accuracy scores for model uncertainty (OOD) and data uncertainty (controlled context) scenarios.

A.4 Linear Regression Analysis

Figure 14 reports all the effects and corresponding coefficients and p-values of the linear regression model described in Subsection 5.3.



```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.58641 -0.10545 -0.06753  0.09504  0.53066 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.0083035  0.0170900   0.486  0.62709    
POSADV      -0.0175029  0.0142610  -1.227  0.21978    
POSNOUN      0.0332515  0.0116023   2.866  0.00418 ** 
POSVERB      0.0687057  0.0098485   6.976 3.61e-12 ***
nMorph      -0.0115582  0.0035480  -3.258  0.00113 ** 
nGT          0.0843417  0.0120718   6.987 3.35e-12 ***
nPD          0.0086235  0.0004789  18.006  < 2e-16 ***
dHypo       -0.0021911  0.0011069  -1.979  0.04785 *  
dSyno       -0.0012973  0.0014049  -0.923  0.35585    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1648 on 3511 degrees of freedom
Multiple R-squared:  0.175,	Adjusted R-squared:  0.1731 
F-statistic:  93.1 on 8 and 3511 DF,  p-value: < 2.2e-16
```

Figure 14: Linear regression model predicting the UE score (SMP) by various effects.