**(一)一种半监督的汉语词义消歧方法**

**Semi-Supervised Method for Chinese World Sense Disambiguation**

**——张春祥，徐志峰，高雪瑶(西南交通大学学报 2019,2(54))**

**一、科学问题**

**1.1 本文所涉及科学问题**

研究解决自然语言处理领域中的一词多义问题

**1.2 同行专家如何解决**

基于语言学模型弥补有效监督消歧方法数据稀疏问题；神经序列学习模型；无指导的跨语言词义消歧算法。

**1.3 本文所解决的问题**

词义消歧准确率提高了0.9%，其性能相比基于支持向量机(SVM)的方法有所提高。

**1.4 本文解决方案效果**

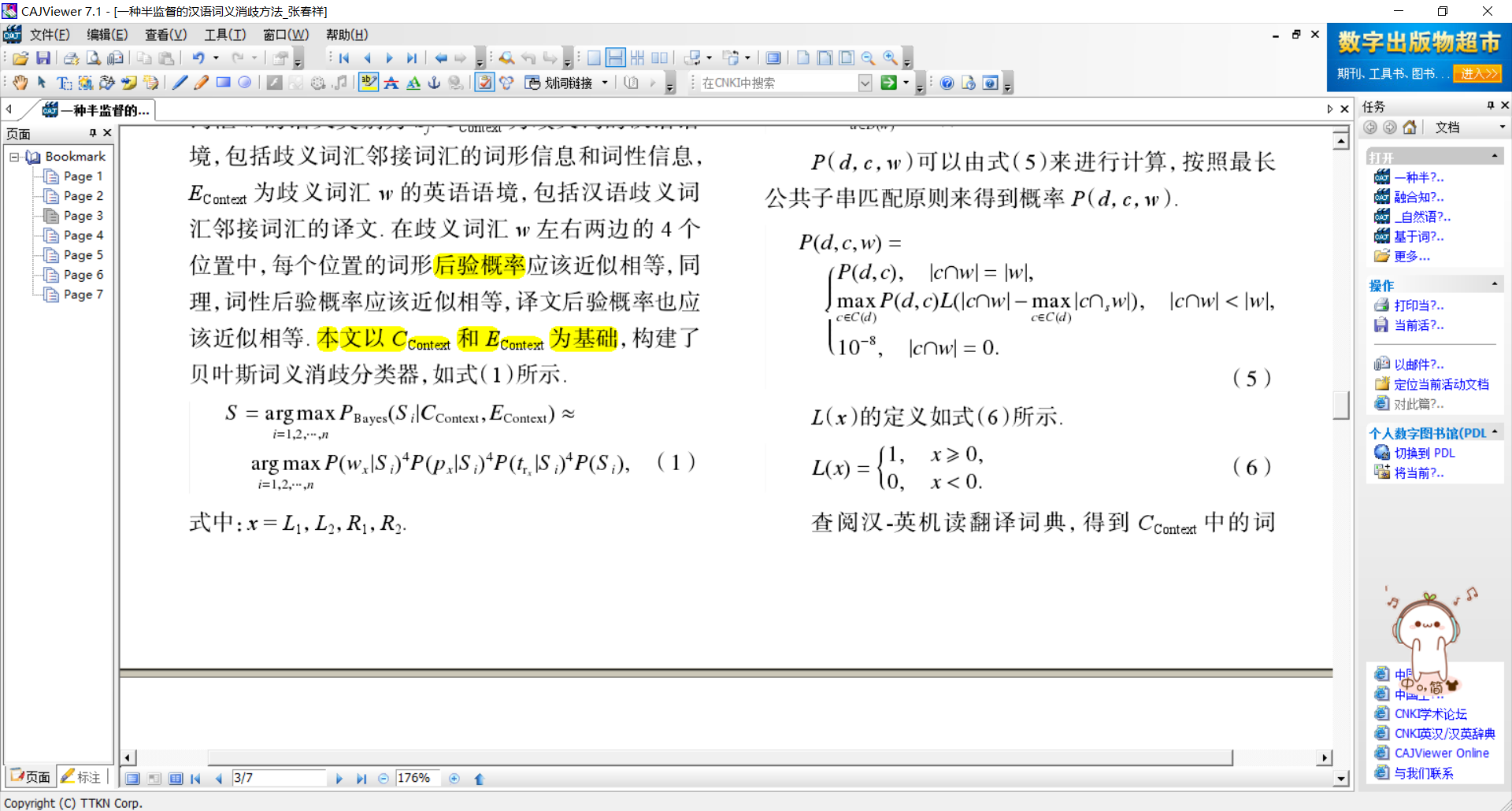
消歧模型相互促进从中挖掘更多语言学知识，词义消歧性能有所提高。

**二、探究内容**

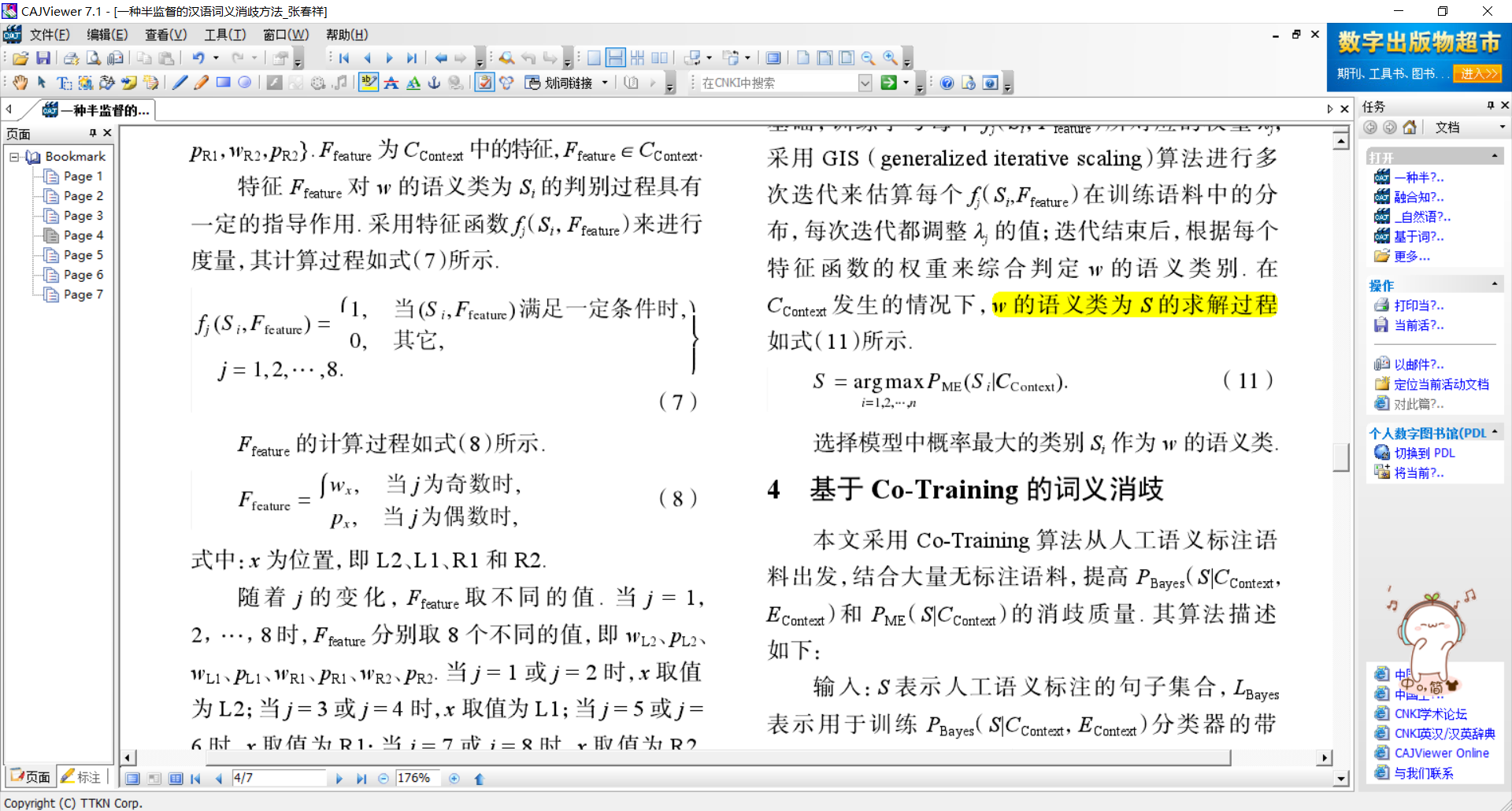
**2.1 理论与方法介绍**

将歧义词汇左右邻接4个单元词汇的词形、词形和译文作为消歧特征；

以汉语和英语语境为基础构建贝叶斯模型：

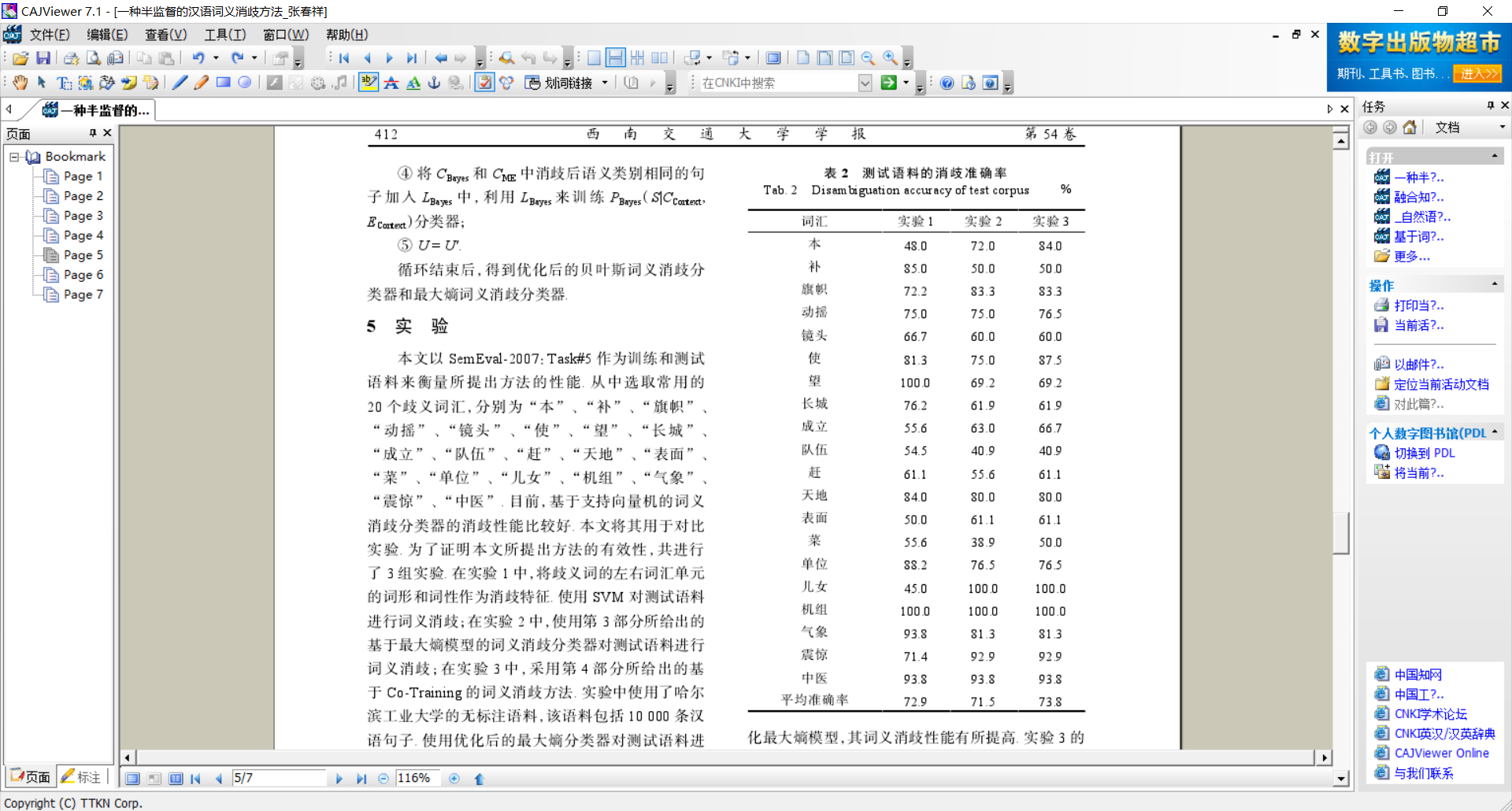
****；

根据特征函数计算其权重构建最大熵模型：

；

采用Co-Traning算法从人工语义标注语料，结合大量无标注语料进行训练进而优化前两模型。

**2.2 验证分析与实验效果**

****

词义消歧了同类模型中SVM具备较好的分类性能，但采用Co-traning算法等优化后的最大熵模型明显比SVM高，主要综合汉英语言学知识来更好地消歧。

**三、论文存在的问题及后续研究重点**

**3.1 论文存在问题**

实验选用的歧义词汇涵盖范围可以再提升，缺少了对优化后的贝叶斯分类器进行实验。

**3.2 后续研究重点**

研究改进算法提高词义消歧准确率；利用更多的语言学知识来提高模型能力。

**四、该问题相关研究成果**

**4.1 相关论文一**

**（1）题目：**Improving selection of synsets from Wordnet for domain-specific word sense disambiguation(改进从WordNet中选择的同义词，用于特定领域词义消歧)

**（2）作者介绍：**Ivan Lopez-Arevlo，Victor J.Sosa-Sosa，Franco Rojas-Lopez，Edgar Tello-Leal

**（3）摘要:** 本文通过选择模糊词的主要意义来描述针对特定领域的词义消歧的方法，主要有四个阶段：（1）辅助语料库生成、（2）相关特征提取、（3）测试特征提取、（4）特征集成。本文主要关注的特定领域语料库的获取结果。

**4.2 相关论文二**

**（1）题目：** 基于领域知识的图模型词义消歧方法

**（2）作者介绍：** 鹿文鹏，黄河燕，吴昊

**（3）摘要：**本方法利用文本领域关联词和句子上下文词构建消歧图，并根据词义领域知识对消歧图进行调整，使用改进的图评分方法对消歧图的各个词义节点的重要度评分。

**4.3 相关论文三**

**（1）题目：**基于语言模型的有监督词义消歧模型优化研究

**（2）作者介绍：**杨陟卓，黄河燕

**（3）摘要：**优化传统的有监督消歧模型，充分利用有监督和语言模共同推到歧义词的词义。可以在训练语料不足的情况下，有效的提高词义消歧效果。

**(二)基于语言模型的有效监督词义消歧模型优化研究**

Supervised WSD Model Optimization Based on Language Model

**——杨陟卓，黄河燕（2014）**

**一、科学问题**

**1.1 本文所涉及科学问题**

缺乏大规模的训练语料，有监督词义消歧方法还不能取得满意的结果，本文研究语言模型的辅助消歧效果。

**1.2 同行专家如何解决**

在有限的标注语料条件下，提高有监督词义消歧效果的有效途径之一是充分利用句子中的各种特征。相关方法有：自动选择词义消歧模板特征方法；采用语义分析技术识别出实体类型再将其语义信息作为特征训练分类器；针对消歧模型进行优化。

**1.3 本文所解决的问题**

利用两种模型的消歧优势，共同推到歧义词词义，可以在训练语料不足情况下，有效地提高词义消歧效果。

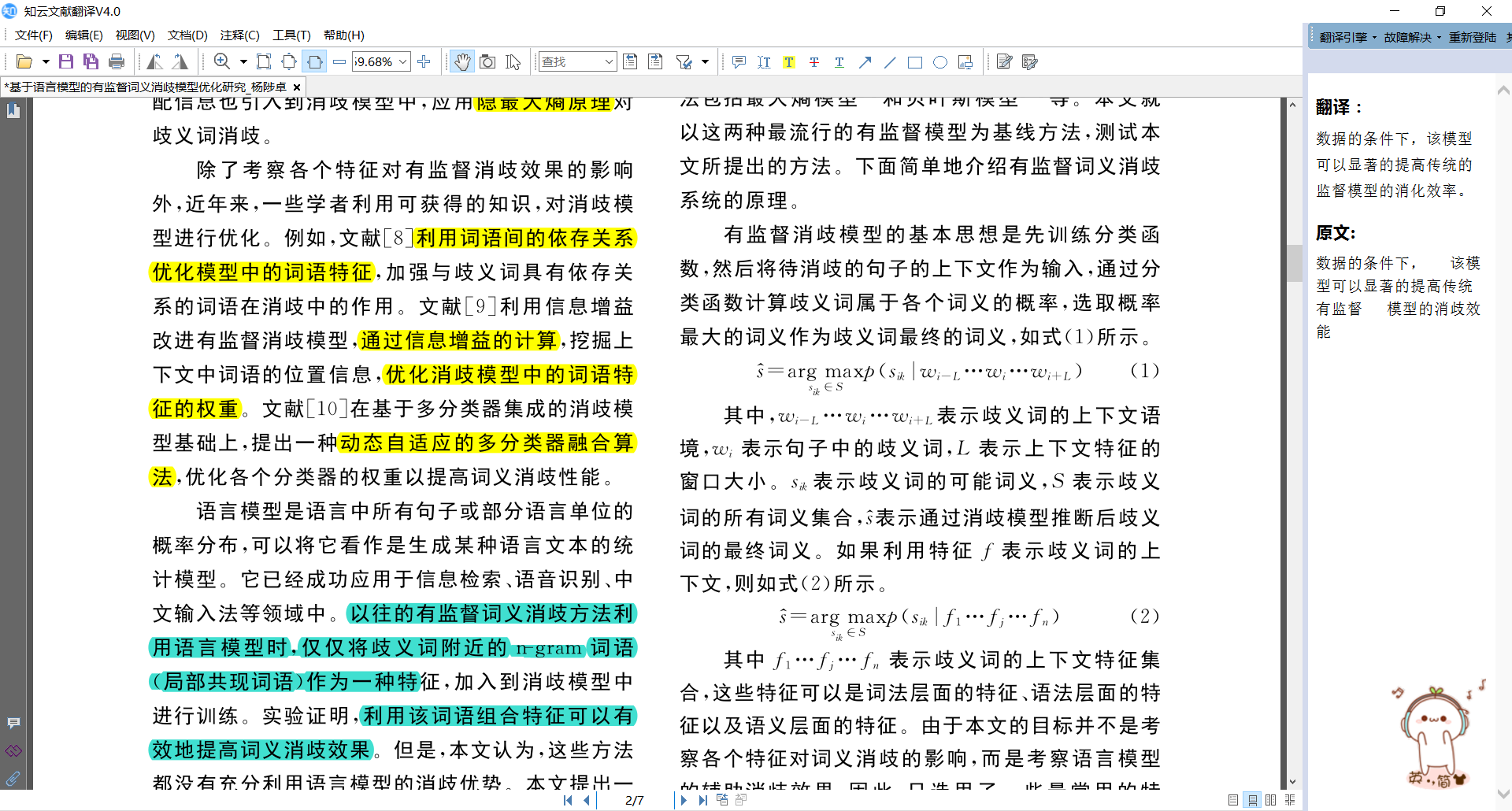
**1.4 本文解决方案效果**

在传统的有监督模型基础上，利用语言模型的词义消歧优势，弥补有监督消歧模型存在的数据稀疏问题，在缺乏训练数据条件下，该模型可以显著的提高传统有监督模型的消歧效能。

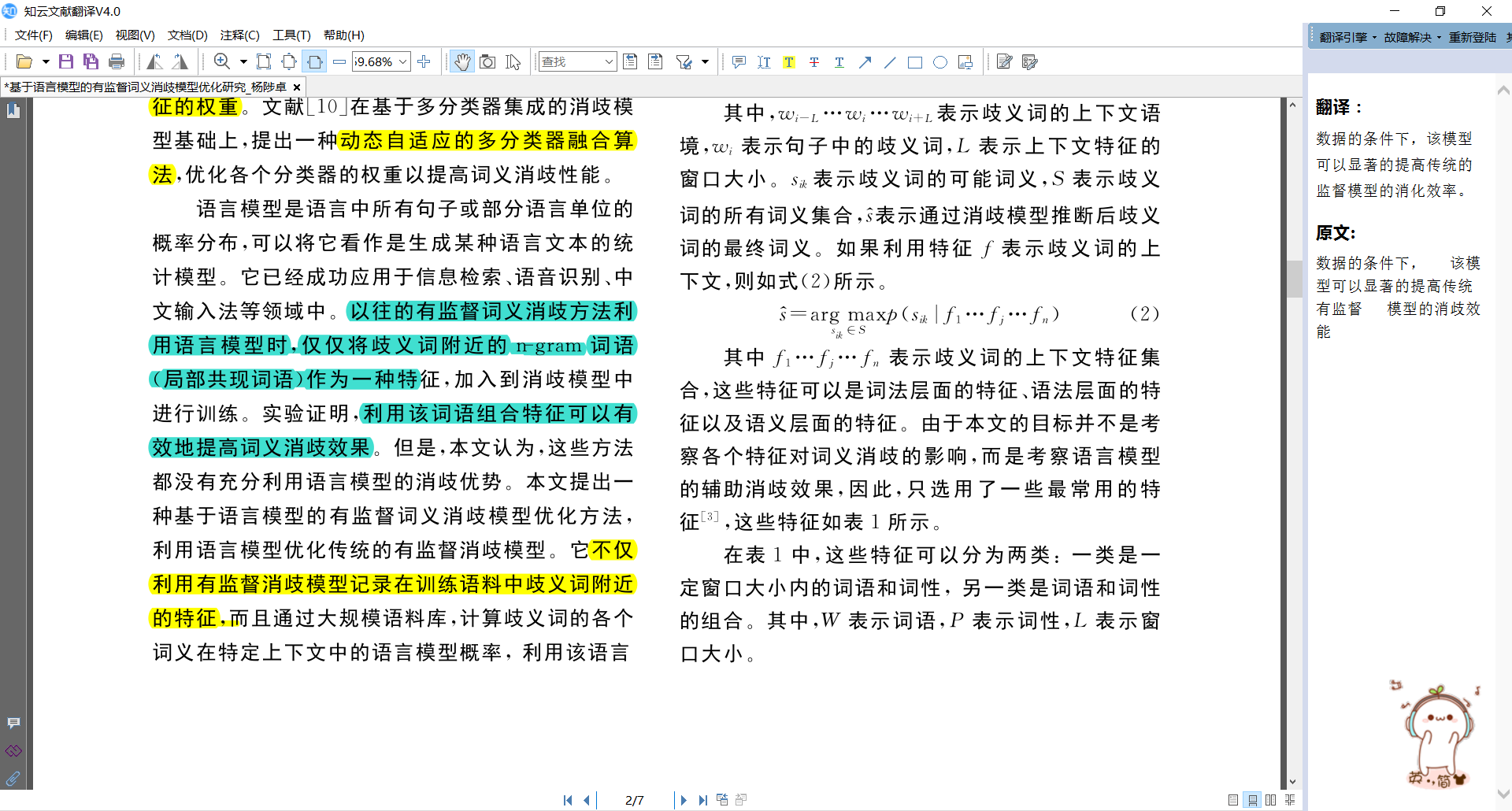
**二、探究内容**

**2.1 理论与方法介绍**

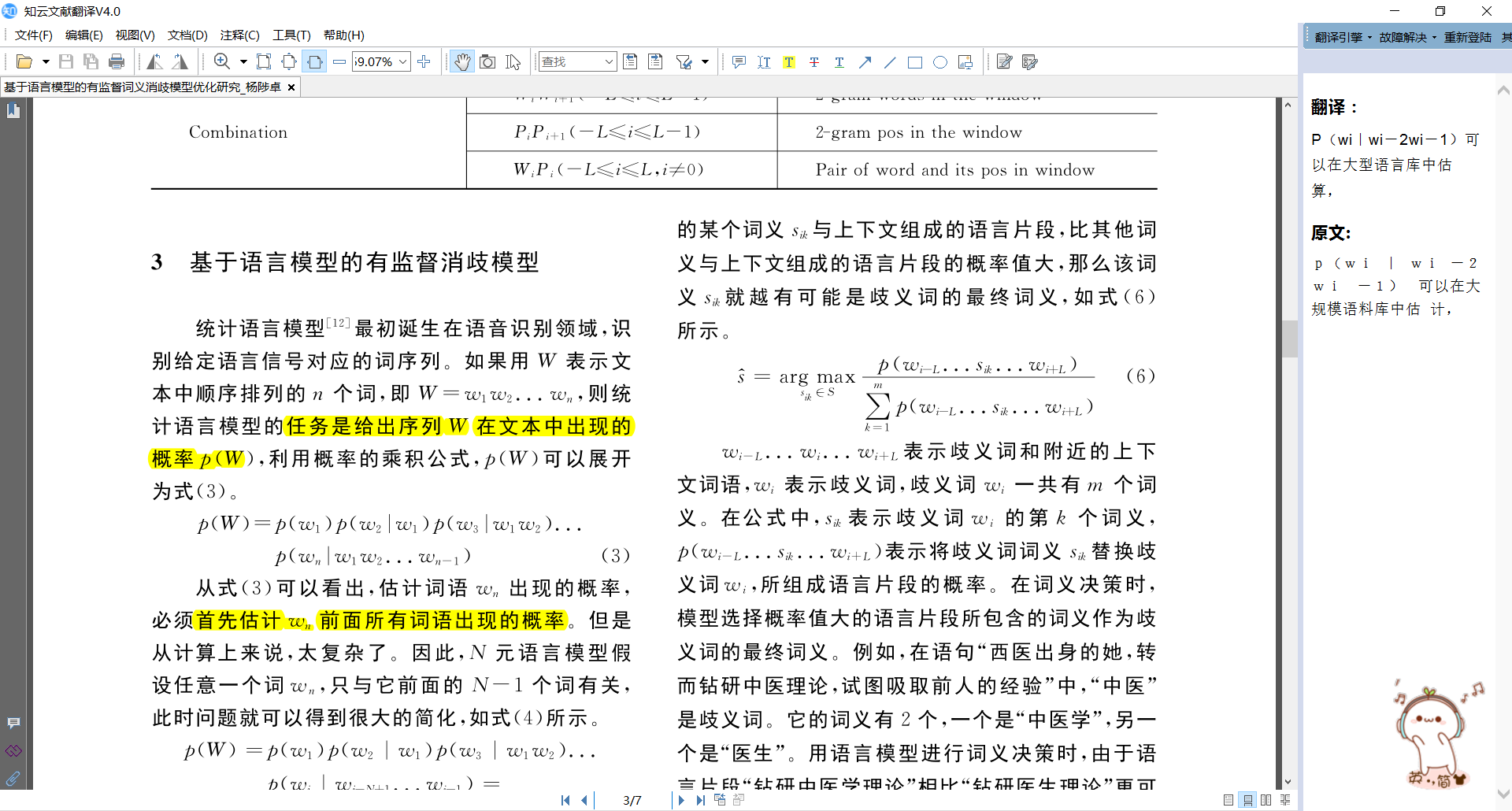
本文采用最大熵模型和贝叶斯模型的有监督消歧模型为基线方法，待消歧句子的上下文输入，用训练过后的分类函数计算歧义词属于各个词义的概率，选取概率最大的词义作为歧义词最终的词义：

****

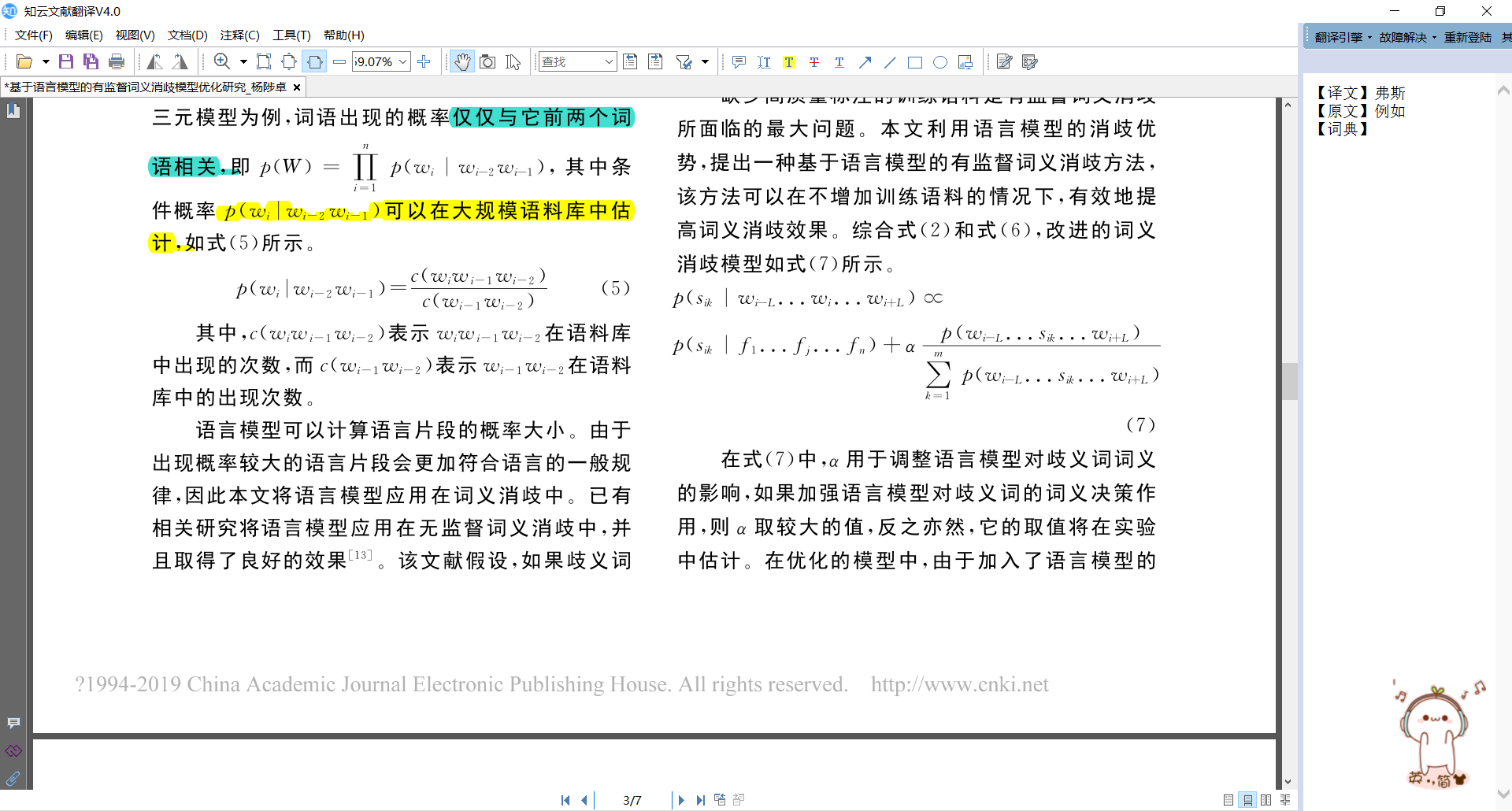
如果利用特征f表示歧义词的上下文：

；

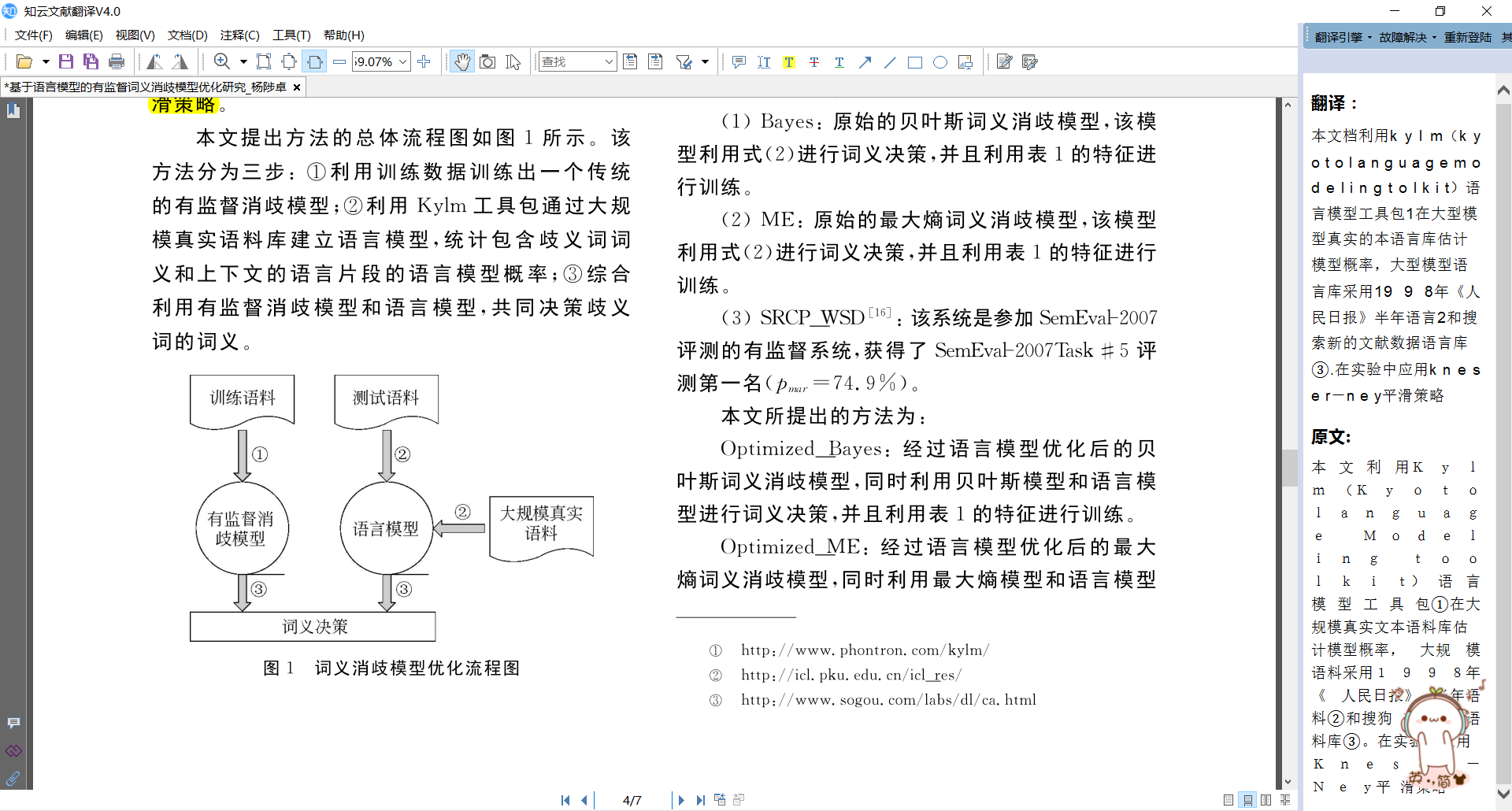
采用语言模型消歧优势，词义与上下文组成语言片段概率越大就越有可能是歧义词最终词义：

；

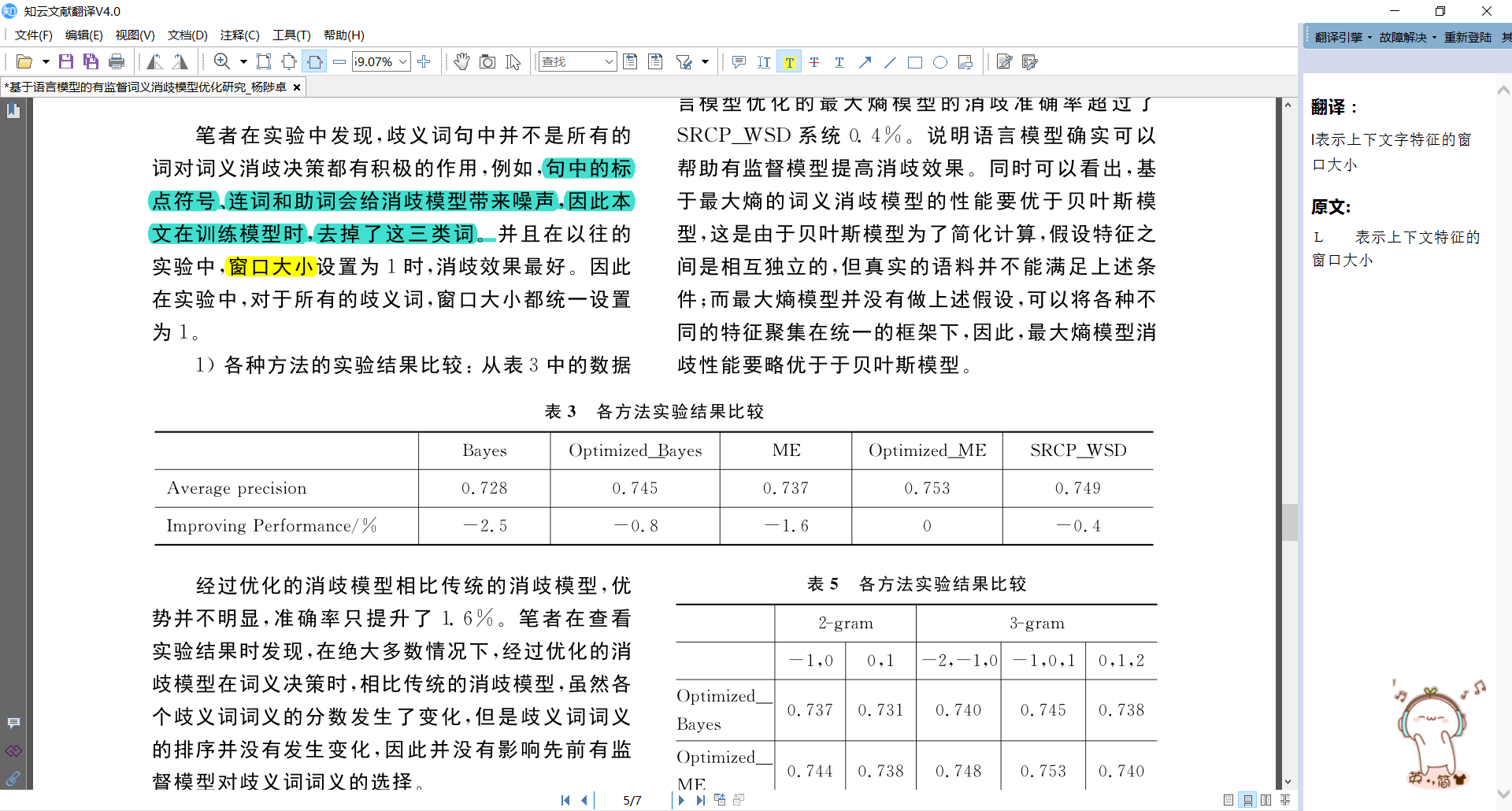
利用语言模型来优化传统有效监督消歧模型：

；

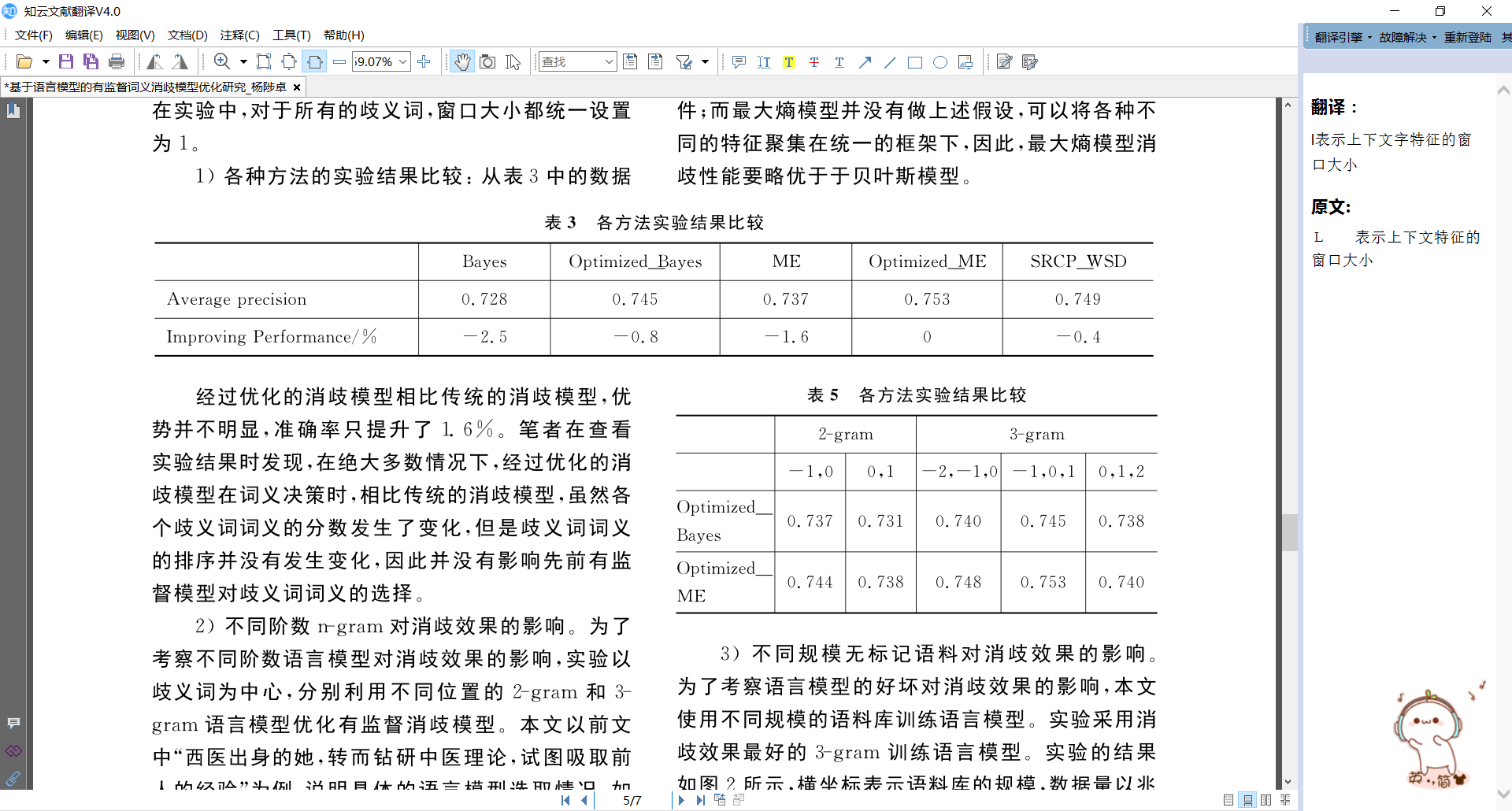
利用Kylm语言模型工具包在1998年《人民日报》半年语料和搜狗新闻数据语料库中估计模型概率，应用Knser-Ney平滑策略。

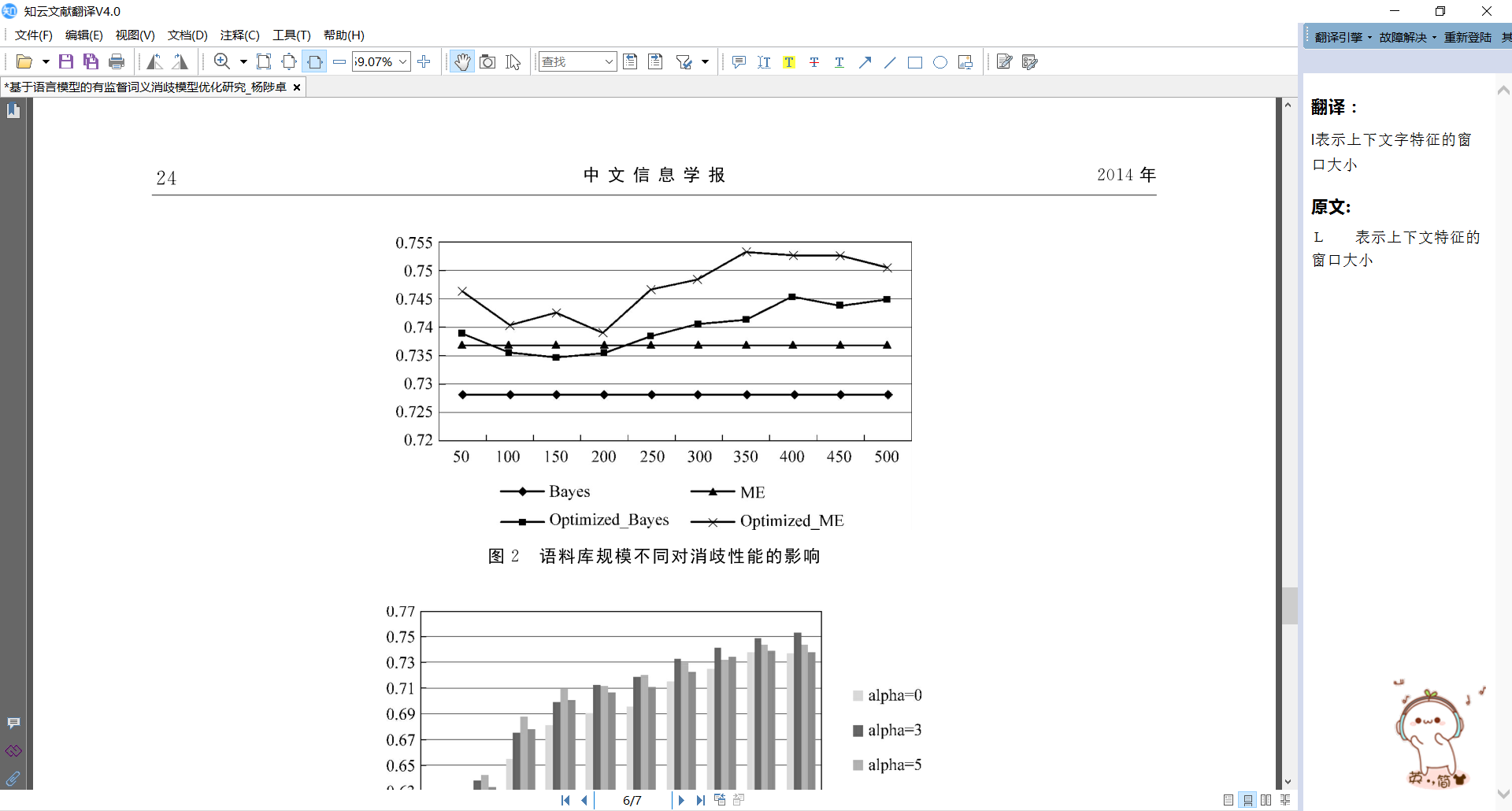


**2.2 验证分析与实验效果**

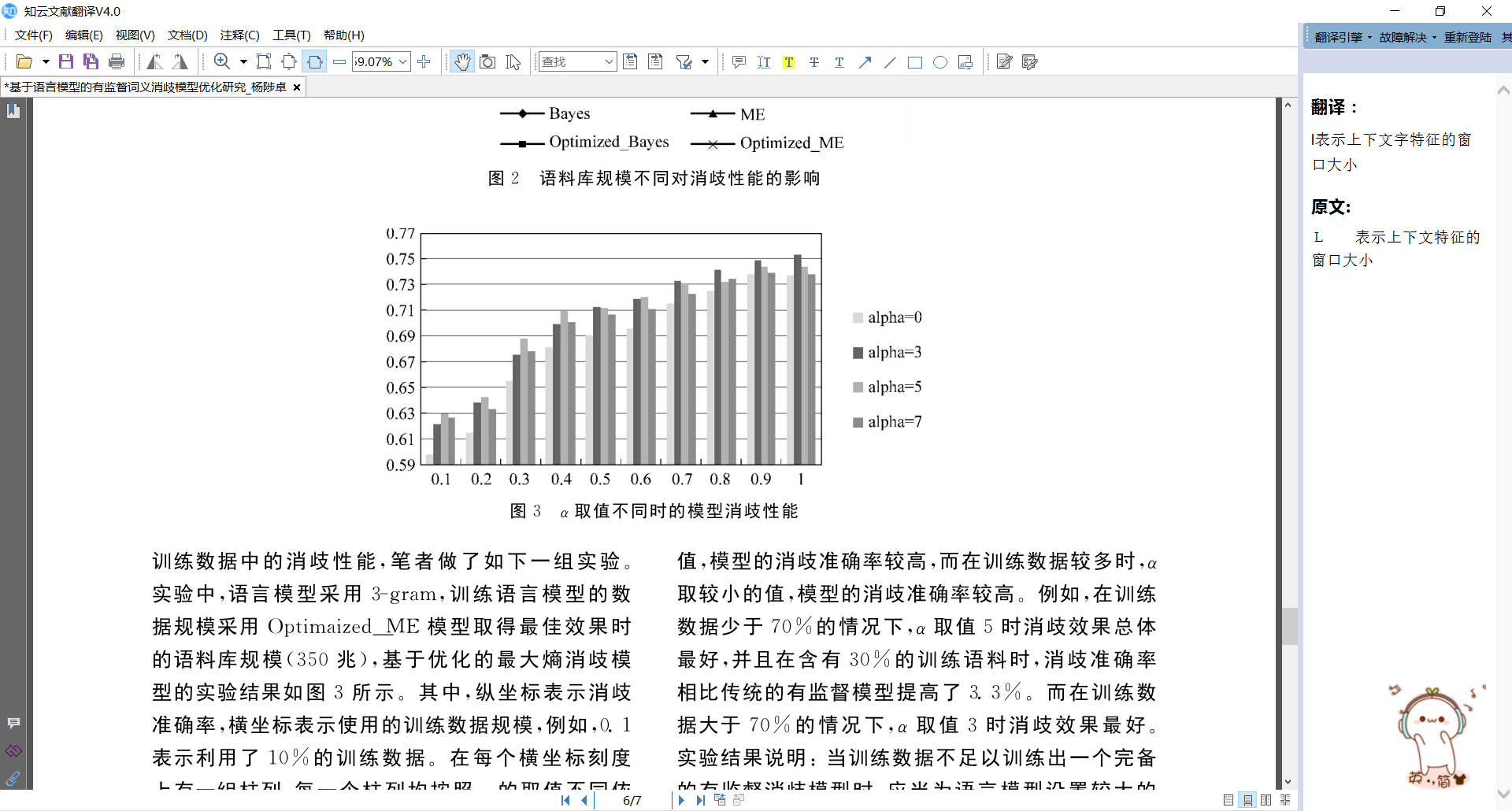


语言模型可以帮助有监督模型提高消歧效果，但经过优化的消歧模型优势并不明显，准确率只提升了1.6%。

****

****

采用3-gram训练语言模型，可看出语言模型对消歧性能的提高有限。



**三、论文存在的问题及后续研究重点**

**3.1 论文存在问题**

语言模型的有监督消歧模型例子可以举歧义多点的歧义词来验证；针对不同阶数n-garm的选取只选了2和3；对于人民日报语料库和搜狗新闻医疗库所训练的语言模型的消歧效果比较不够严谨。

**3.2 后续研究重点**

深入分析大规模中文词义消歧数据集的特征，挖掘更多可用的词义消歧知识；

研究如何优化其他有监督词义消歧模型。

**四、该问题相关研究成果**

**4.1 相关论文一**

**（1）题目：**基于领域知识的图模型词义消歧方法

**（2）作者介绍：**鹿文鹏，黄河燕，吴昊

**（3）摘要:** 本方法利用文本领域关联词和句子上下文词构建消歧图，并根据词义领域知识对消歧图进行调整，使用改进的图评分方法对消歧图的各个词义节点的重要度评分。

**4.2 相关论文二**

**（1）题目：**Improving selection of synsets from Wordnet for domain-specific word sense disambiguation(改进从WordNet中选择的同义词，用于特定领域词义消歧)

**（2）作者介绍：**Ivan Lopez-Arevlo，Victor J.Sosa-Sosa，Franco Rojas-Lopez，Edgar Tello-Leal

**（3）摘要：**本文通过选择模糊词的主要意义来描述针对特定领域的词义消歧的方法，主要有四个阶段：（1）辅助语料库生成、（2）相关特征提取、（3）测试特征提取、（4）特征集成。本文主要关注的特定领域语料库的获取结果。

**4.3 相关论文三**

**（1）题目：**一种半监督的汉语词义消歧方法

**（2）作者介绍：**张春祥，徐志峰，高雪瑶

**（3）摘要：**利用多种语言学知识和词义消歧模型的半监督消歧办法，以歧义词左右邻接词单元的词形、词性和译文作为消歧特征来构建贝叶斯词义分类器，并以词形和词性作为消歧特征来构建最大熵词义分类器，利用Co-Training算法并结合大量无标注语料来优化模型。