南京大学

研 究 生 毕 业 论 文

（申 请 硕 士 学 位）

论　文　题　目　句子级与片段级潜台词分析研究

作　者　姓　名　严骅

专　业　名　称　计算机科学与技术

研　究　方　向　自然语言处理、潜台词分析、情感分析

指　导　教　师　申富饶 教授

2022 年 **5** 月 **22** 日

学　　号：**MG1937028**

论文答辩日期：**2022** 年 **5** 月 **17** 日

指 导 教 师：　　（签字）

The Research of Sentence-level And Segment-Level Subtext Analysis

by

Yan Hua

Supervised by

Professor Shen Fu-Rao

A dissertation submitted to
the graduate school of Nanjing University
in partial fulfilment of the requirements for the degree of
MASTER
in
Computer Science and Technology

School of Artifical Intelligence
Nanjing University

May 20, 2022

南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目：句子级与片段级潜台词分析研究

计算机科学与技术 专业 2019 级硕士生姓名：严骅

指导教师（姓名、职称）：申富饶 教授

摘　要

潜台词是“某一话语的背后，所隐藏着的那些没有直接、明白表达出来的意思”，它要求读者对文字所要表达的内容有更深层次地理解。潜台词不仅存在于中文，在不同的语言文化中都有使用潜台词的现象且对潜台词的认知保持一致。在目前为止的研究工作中，却鲜有关注和研究潜台词的。然而，潜台词不仅作为日常交流的意图表现方式被广泛使用，在文学作品中的使用也是屡见不鲜。若机器对于潜台词的理解得以实现，对于通用对话系统和领域对话系统都将有突破性地改变：让机器更加智能，让对话更加流畅。不仅如此，让机器读懂潜台词可以在教育领域中对阅读理解相关的教学活动提供极大的助力，如辅助出题和判罚等等。可见，潜台词有着其重要意义。鉴于此，本文提出了潜台词分析相关问题，包括潜台词检测、潜台词片段识别和潜台词还原，并从社交网站中搜集相关数据，对其进行标注与分析。本文的主要内容均基于此进行展开。

本文的主要内容如下：

1. 本文提出了潜台词分析，包括潜台词检测、潜台词片段识别和潜台词还原，并给其下了完整的定义，将各个子任务进行区分。
2. 本文为潜台词分析任务构建了相关数据集，命名为 CSD-数据集。该数据集的原始数据来源于知名社交网站。本文对原始数据进行标注和分析。过程中，本文提出 TAE score 这一新的衡量指标，显式地建模了一致性、随机性和可靠性之间的关系，并且在开源的 Amazon 数据集上进行了验证。
3. 为解决潜台词检测问题，本文提出 SASICM 算法，并将之与相关的模型进行比较和分析。于 SASICM 算法中，本文建模了目标语句 (本文后续内容所对应的 Comment) 和完整上下文之间的关系。SASICM 将目标语句映射到语义子空间中，学习目标语句在有上下文和无上下文的情况下的语义特征

表示，通过语义特征表示判断是否语义一致，如果语义不一致则含有潜台词，否则不含有潜台词或无法判断。同时，为了解决 Self-Attention 中注意力值区分度不明显的问题，本文提出强化注意力 (Strengthen Attention)，在潜台词检测问题上的 F_1 指标取得了 1.5% 的提升。

4. 为解决潜台词片段识别问题，本文结合嵌套命名实体识别任务共同分析，提取二者共同的特征并提出中心度的概念。而后，本文结合潜台词片段识别问题的独有特征为其设计了基于中心度的 CBM 模型，并在 CSD-数据上进行实验，分析了不同模型在该数据集上的效果与不足，明确了未来工作中可以继续改进的方向。同时，本文还将中心度应用于嵌套命名实体任务设计了 CBLM 模型，与当前最好的模型进行比较，取得了不错的效果：在相同的运行速度下，CBLM 的效果最好；实验效果相近的情况下，CBLM 的运行速度最快，且不需要复杂的预处理。
5. 本文将所提出的潜台词检测、潜台词片段识别集合到一个演示系统中，供用户进行使用。系统中，本文还聚合了数据搜集和标注的功能。从演示系统的效果而言，本文算法对于潜台词检测与潜台词片段识别均有效用。

关键词： 潜台词检测，潜台词片段识别，CSD-数据集，强化注意力，中心度

南京大学研究生毕业论文英文摘要首页用纸

THESIS: The Research of Sentence-level And Segment-Level Subtext Analysis

SPECIALIZATION: Computer Science and Technology

POSTGRADUATE: Yan Hua

MENTOR: Professor Shen Fu-Rao

Abstract

Subtext refers to “the hidden meaning that is not directly and clearly expressed behind a certain utterance”, which requires readers to have a deeper understanding of the content to be expressed by words. Subtext exists not only in Chinese, but also in different languages and cultures. Subtext is rarely mentioned in existing works. However, subtext is a popular way of experssion not only in our daily communication, but also in literary works. If the machine can understand the subtext behind the literal meaning, it will be a great breakthrough for both the general dialogue system and the domain-specific dialogue system：the machine can be real intelligent and the conversation between human beings and robots can be more fluent. Moreover, understanding subtext for machines can help to the educational activities a lot, such as helping teachers in creating questions and judging scores and so on. Therefore, subtext is significant. In view of this, subtext analysis and its sub-tasks are proposed by this paper, including subtext judgment, subtextual segment recognition and subtext reduction. This paper collects relevant data from social networking sites, and we annotate and make a detailed analysis of these data. The main content of this article is described as follow.

1. This paper proposes subtext analysis, including subtext detection, subtextual segment recognition and subtext reduction, and makes clear definitions of them and make clear differentiations for them.
2. This paper constructs a dataset, termed as CSD-dataset, for subtext analysis. The source data of CSD-dataset comes from well-known social networking sites. We annotated the source data and made a detailed data analysis. In analysing process, TAE Score, a new evaluation metrics, is proposed, which explicitly models the

relationship between agreement, randomness and reliability. Moreover, we verified the reliability of TAE score in the open-sourcing Amazon dataset.

3. In order to solve the problem of subtext detection, SASICM algorithm is proposed and compared with other models. In SASICM algorithm, this paper models the relationship between the target statement (comment is the corresponding content from hence) and the full context. The subtext is determined by judging whether they agree with each other that the semantic representations of the target statement with or without context. At the same time, in order to solve the problem of poor differentiation of attention value caused by self-attention, this paper introduces Strengthen Attention, and the F_1 score of subtext detection has been improved by 1.5%.
4. In order to solve the problem of subtextual segment recognition, this paper analyzes the tasks of nested named entities together, and puts forward the concept of centerness after analysing their common features. Then, combining with the unique characteristics of subtextual segment recognition, a CBM model based on centerness is constructed, and experiments are carried out on the CSD dataset. This paper analyzes the effects and shortcomings of different models to clarify what can be further improved in future work. At the same time, this paper also applies centerness to the task of nested named entity recognition by designing the CBLM model. Compared with the state-of-the-art models, it has achieved a well result: at the equivalent running speed, CBLM has the best effect; in the equivalent effect, CBLM has the fastest reasoning speed and it does not need complicated preprocessing.
5. In this paper, the subtext detection and subtextual segment recognition are integrated into a demonstration system for users. Meanwhile, this system aggregates the functions of data collection and annotation. From the performances displayed in the demonstration system, it can be seen that the models (SASICM and CBM) are both useful and effective.

keywords: Subtext Detection, Subtextual Segment Recognition, CSD-Dataset, Strengthen Attention, Centerness

目　　录

插图清单

附表清单

第一章　绪论

1.1　潜台词的研究背景及其研究意义

1.1.1　潜台词的研究背景

潜台词是文学作品和日常交流中常用的表达方式。它要求读者对文字所要表达的内容有更深层次地理解。潜台词不仅存在于中文，不同的语言中都有使用潜台词的现象。不同语言中对潜台词都有定义，且具有高度的一致性，可归结为“某一话语的背后，所隐藏着的那些没有直接、明白表达出来的意思”，简言之，当一句话具有两层含义，一层文字表面意思 (Literal Meaning)，一层话者的实际含义 (Figurative Meaning)，且两层意思的指向不同时，该句具有潜台词。然而，无论是工程使用中或者是学术研究中，对于潜台词的关注或者研究都极少。学术研究和工程落地中关注的重点多是文本分析中的情感分析、比喻分析、反讽分析、情绪分析等领域。鉴于此，本文对潜台词进行分析研究，力求填补这方面的空白。

本文将潜台词分析分成三个子任务：潜台词检测（Subtext Detection），潜台词片段识别 (Subtextual Segment Recognition) 和潜台词还原 (Subtext Recovery)。为了更好地说明这三者，本文对符号表示先行约定，记 $S = \{s_1, s_2, \cdots, s_n\}$ 为需要分析的目标序列，此处 S 本身是语义完整的[①]，$C = \{c_1, c_2, \cdots, c_m\}$ 为目标序列 S 对应的上下文序列或者是背景知识序列，C 本身也是语义完整的，$f(x)$ 为语义特征抽取函数，C_o 为常识知识或者是世界知识。

- 潜台词检测：潜台词检测是一个句子级别的分类任务问题。潜台词检测的输入为 $[S;C]$，目标是判断目标序列 S 在有上下文和无上下文的情况下，它实际要表达的意思与其字面意思是否不一致。以“你是个好人”为例，它的字面意思就是简单的夸赞一个人“好”，但是在上下文或者背景提要中“我喜欢你，你可以做我女朋友吗？”，这句话的意思就变成了“我不喜

[①]本文认为语义完整为当前句子可以独立成句，在没有上下文或者更多说明的情况下，依旧可以完整表达话者的意图。

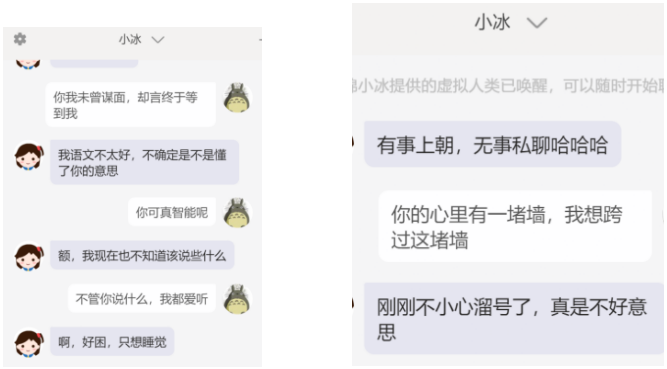
欢你”。显然，这句话，在上下文的条件下，它实际要表达的意思与字面上要表达的含义变得不一致，本文称目标序列 S（“你是个好人”）在上下文 C（“我喜欢你，你可以做我女朋友吗？”）下，含有潜台词。用数学的形式来进行描述为：当 $f(S \mid C_o) \neq f(S \mid C, C_o)$ 时，S 具有潜台词。在有足够多的训练语料时, 世界知识或者常识知识 C_o 可以认为已经被预训练的语言模型，如 Word2Vec[7]、GloVe[8] 或者 BERT[9] 等，编码进文本的分布式表示 (Embedding) 中。因此，上式可以简写为：$f(S) \neq f(S \mid C)$。我们将判断 $f(S)$ 与 $f(S \mid C)$ 是否相等这个任务定义为潜台词检测，该问题的目标为学习一个判断函数 $I(\cdot)$, 当 $I(f(S), f(S \mid C)) = 1$ 时，S 具有潜台词，当 $I(f(S), f(S \mid C)) = 0$ 时，无法判断是否具有潜台词，$I(f(S), f(S \mid C)) = -1$ 时，S 不具有潜台词。

- 潜台词片段识别：潜台词片段识别是一个片段级 (Segment-Level) 的分类问题，若 S 具有潜台词，对 S 进一步分析，识别在 S 中含有潜台词的原文内容（文本片段），即给定一个具有潜台词的目标序列 S，识别一段子序列 $S_{i:j} = \{s_i, s_{i+1}, \cdots, s_j\}$，使得 $f(S \mid C) = f(S_{i:j} \mid C)$。以上文中 “你是一个好人” 为例，该处需要识别出 “好人” 这一个文本片段。
- 潜台词还原：潜台词还原是一个类似于机器翻译的问题。若 $S_{i:j}$ 是一个潜台词片段，在给定上下文和完整的目标序列条件下，生成一段子序列 $O_{i:j} = \{o_i, o_{i+1}, \cdots, o_j\}$，使得 $f(S \mid C) = f(S_{S_{i:j} -> O_{i:j}})$，其中 $S_{S_{i:j} -> O_{i:j}}$ 表示将目标序列 S 中的子序列 $S_{i:j}$ 替换为序列 $O_{i:j}$。为了简化潜台词还原的复杂性，我们将这个过程拆分成两个部分。首先，如果目标序列中具有修饰性描述，则用普通描述替换这个修饰性描述。如将比喻的喻体使用本体进行替换，或将反讽的表达方式使用普通的负面表达进行替换。其次，推测深层次含义，并且生成连贯的语句。本文的研究内容聚焦于前两个问题，对潜台词还原内容的探讨留待后续进行。

1.1.2 潜台词的研究意义

对话系统是自然语言处理领域的一类典型应用，其大体可以分为通用对话系统 (Genera Dialog System)[10] 和领域对话系统 (Domain-specific Dialog System)[11], 二者皆受到学术界和工业界的研究者的广泛关注。潜台词对于通用的对话系统具有更重要的意义。对于通用的对话系统而言，如微软小冰[①]，

[①]微软的一款对话系统



(a) 对话示例 1 (b) 对话示例 2

图 1-1: 微软小冰的对话示例

它所能处理的内容也大多为简单直接的对话，即不具有潜台词的对话。从图1-1中所示内容可得，当我们所说的内容比较委婉甚至是带有冷嘲热讽时，其不能够很好的对所述内容做出恰当地回应。若机器具有分析潜台词的能力则可使对话系统更加智能。

机器翻译[12] 也为当前的热点研究内容。机器翻译的主流做法为设计一个从源语言到目标语言的端到端模型。此做法固然合理但依旧有其无法解决的问题。其中重要的一点为源语言与目标语言的习惯可能不一致，这将导致的问题为在源语言中委婉的习惯表达方式，通过按字面进行翻译到目标语言，难以保留其原始意义。在目前主流的机器翻译解决方法中，该类问题能否避免完全取决于语料中是否有正确的标准翻译，而这往往难以实现。若在机器翻译中加入潜台词分析，则可以在一定程度上避免该问题。潜台词分析使得机器可以直接获取委婉的表达方式所要表达的意思，从具体含义入手进行翻译，则可以保证在其意思不出错前提下进行翻译且保证了让不同语言环境中的人能够看懂。

除此之外，潜台词在教育方面，也有其应用价值。如在阅读理解相关问题上，出题者可以主要关注在题目的合理性和可行性。比如，在鉴赏古典诗词领域，可以使用潜台词分析的技术，对其进行自动化分析，给出候选答案，由出题者进行筛选和复检，这可以帮助出题者快速地出完相关试题。

同时潜台词分析对于舆情分析以及相关的客诉问题也有其意义。在电商领域和一些客户反馈调查领域，难以要求用户结构化输入，又兼具有用户表达方式多种多样的问题。类似的现象在社交网站的舆情分析中也屡见不鲜。当用户喜欢非直接表达时，以“这产品我不得不说好，前天买的，今天就可以换新的了。”为例，该句为典型的非直接表达，比较委婉。它的双重否定暗含的内容

是“这个产品质量非常差劲”，若是机器不能正确理解类似的表达方式，就会对客户的情绪、态度等造成误判。而潜台词分析可以让机器更准确地理解使用者所表达的真实含义。

1.2 潜台词研究的现状及其难点分析

1.2.1 研究现状

如1.1.1中所述，调研结果显示在本文之前，没有其它相关的研究者或者研究工作对潜台词分析问题提出过相关解决方案。因而，本文仅能从相近的工作进行参考。根据本文在1.1.1中对潜台词分析子问题的介绍，潜台词检测、潜台词片段识别可以分别抽象成：隐含语义的句子级别文本分类和片段级别或者短语级别的文本分类问题。关于句子级别的文本分类任务和片段级别或短语级别的文本分类任务，有大量的研究者对其研究，并且提出了一系列的方法。本文以此为基础，在潜台词分析的问题上设计相关的解决方案，并将其应用于其余任务之中。具体内容，于后文进行论述。

1.2.2 难点

由于潜台词的任务于本文中首次提出，其所面临的问题主要有以下几点：

- 定义 (Well-definition): 潜台词是一个语言上的常见现象，使用语言来描述什么是潜台词简单而抽象。本文尝试使用数学的形式对其进行描述，仍不能保证其完备且没有歧义。
- 训练资料 (Corpus): 从调研结果上显示，潜台词所必需的训练资料也是没有的，即本文的研究需自行构建相关的训练资料。构建训练资料需解决构建什么类型的语料、怎么避免主观性、怎么衡量语料是否合理等问题。
- 模型 (Model): 关于如何挖掘隐含意义，是研究者们致力于研究的问题。本文基于神经网络的方法对潜台词检测和潜台词片段识别两个问题分别设计了两个模型。对于传统的文本分类问题，他们的研究目标和上下文的语义通常都具有一致性，因此他们从本质上而言是对特征的提纯。然而在潜台词检测上，除了对特征的提纯以外，更重要的是去建模目标语句在有无上下文条件下是否会有语义差别。同时，研究现状表明如何提取一个隐含的语义还未曾有一个良好的解决方案。本文所设计的模型在同一个语义子空

间中对语义进行特征表示，然后使用一个分类模型对他们进行判别。关于潜台词片段识别，目前相关的所有算法，存在的问题为可以概述为两个，其一，它们需要枚举的时间复杂度过高，为 $O(n^2)$ 的时间复杂度；其二，它们所需要的预处理都过于复杂，不利于实现与设计。为此本文提出了一种复杂度较低的解决方案，直观且无需特殊设计即可实现的算法。同时该方案在相似的问题上也取得了很好的效果。

1.3 本文的贡献

本文为潜台词分析构建了相关的数据集，并且提出了在本文所述场景下的一个衡量方法并且与传统方法进行比较。在构建好的数据集上，本文针对潜台词检测设计了 **S**trengthen **A**ttention Based **S**equence and **I**ntra-Attention **C**onfused **M**ulti-Task Model (SASICM) 模型，以及针对潜台词片段识别设计了 **C**enterness **B**ased **M**ulti-task Content Recognition Model (CBM)。除此之外，我们将潜台词相关问题及其算法进行整合，构建成了一个演示系统供给演示及测试，以此验证其有效性。其中本文的主要内容如下：

- 本文构建了潜台词问题的相关数据集 (Chinese-Subtext Dataset, CSD)。我们从相关的网站上，通过网络爬虫的方式对公开的数据进行抓取。经过匿名化处理之后，对数据进行两轮非完全独立标注。同时，对标注的数据集采用两种方式相结合的方法进行衡量：Coherence Kappa Score[13–15] 和 Two-round Annotation Evaluation Metrics (TAE)，其中 TAE 为本文在本文场景下所提的方法，并且通过一些简单的变化，就可以适用于传统的标注场景。TAE 避免了 Kappa 在极端不平衡条件下对数据质量衡量不准确的问题[16,17]。
- 本文基于构建的 CSD 数据集，在潜台词检测问题上，我们设计了 SASICM 模型。在 SASICM 模型中，我们改进了 self-attention 模型即上文所述的 Strengthen Attention，该模块缓解了 self-attention 在本文任务中句子的注意力只集中与自身的情况。同时，我们通过数据分析，对比了采用多任务模型和不采用多任务模型之间的差别。
- 本文基于潜台词片段识别和嵌套命名实体这类任务的共同结构特性，创新性地提出了中心度 (centerness) 的概念，并且根据潜台词片段识别和嵌套命名实体的各自特性分别设计了两种基于中心度的解决方案。本文基于 CSD 数据集，为潜台词片段识别设计了基于中心度的多任务内容片段识别模型

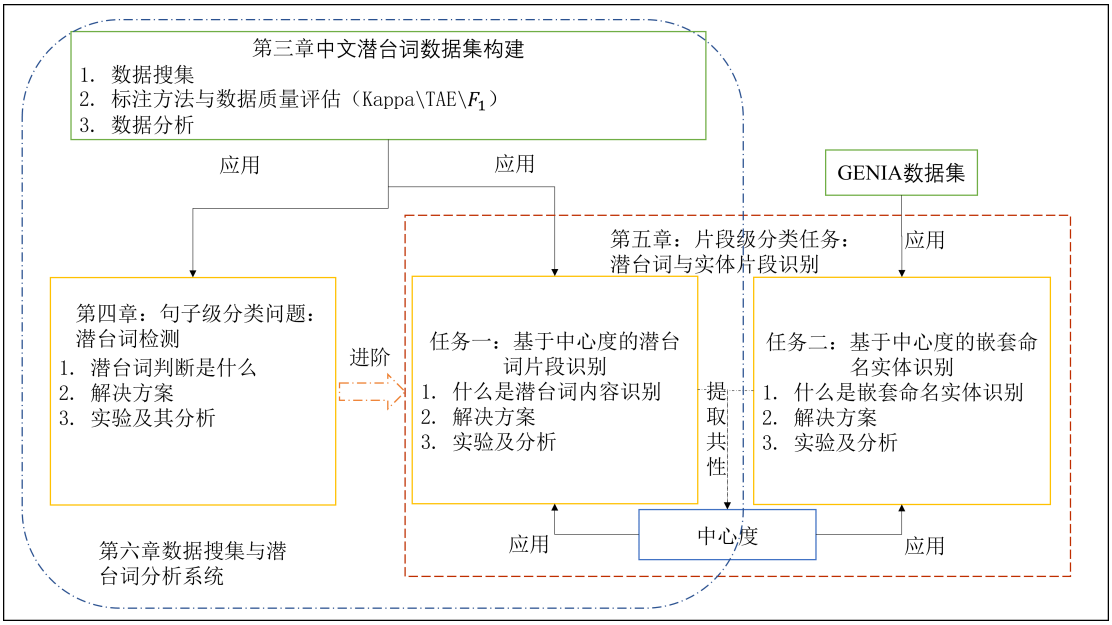


图 1-2: 三 ~ 六章关系图

(CBM)，对普通的潜台词和带有反讽、比喻等的潜台词进行内容片段识别。本文基于开源的数据集 GENIA，为嵌套命名实体识别设计了基于中心度的层级模型 (CBLM)，对嵌套类型的实体进行识别。两个任务上都可以得出，基于中心度的算法可以在效果和推理速度上取得平衡：在相同推理速度下，有最好的效果；在相近的效果上，有最快的推理速度。

- 本文将潜台词检测和潜台词片段识别集成到一个演示系统之中。在该系统主要分析功能完全应用本文算法，给定上下文条件下，本系统可以自动化分析目标语句是否属于潜台词，并且分析出潜台词或者比喻、反讽等所对应的原文内容。

1.4 论文纲要

本文内容围绕潜台词相关任务进行展开，包括潜台词检测和潜台词片段识别，分别提出了对应的算法，一个是带有 strengthen attention 的 SASICM 模型，另一个基于中心度的 CBM 算法。并且我们将两部分进行整合形成一个演示系统，供用户使用。除此之外，本文还将中心度的方法应用在嵌套命名实体之中，亦取得了良好的成绩。本文一共分为七章：第一章节为绪论，介绍了本文的研究背景、意义，当前问题的研究现状以及难点；第二章介绍了数据集构建、句子级别的文本分类问题和片段级别分类问题相关的研究工作；第三章主

要介绍本文所研究问题的数据集及其数据分析；第四章主要介绍本文提出的潜台词判别解决方案——SASICM——的设计细节及其实验分析；第五章主要介绍了基于中心度的算法及其在潜台词片段识别上的应用和在嵌套命名实体识别上的应用，并进行了相应地实验分析；第六章主要介绍如何本文提出的问题及其算法整合到一个演示系统当中；第七章总结全文，并对未来工作进行展望。为了更好的展示三～六章之间的关系，可参见图1-2。

第二章　相关工作

本文聚焦的内容为潜台词相关数据集的构建、潜台词检测以及潜台词片段识别三个方面。关于这三个方向，虽然没有直接相关的参考文献，但是从数据集通用的标注和衡量方法、抽象的文本分类建模和文本序列化标注建模的角度，具有可参考借鉴的一些研究工作。本章将这些研究工作按照以下分类进行介绍：数据集、句子级别的文本分类、嵌套结构的序列标注。

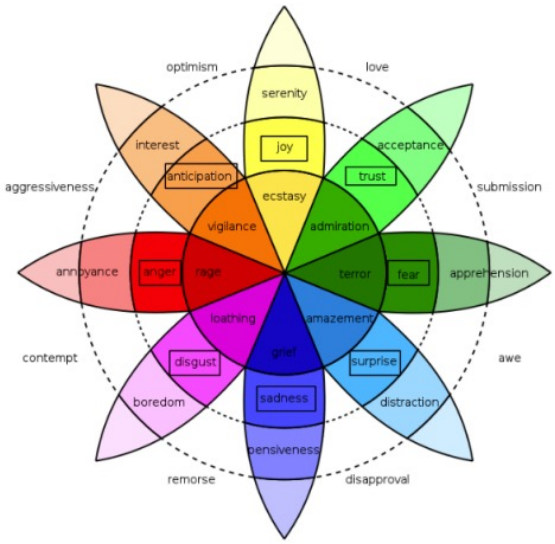


图 2-1: Plutchik 的情绪轮式模型[1]

2.1　数据集

Mishra 等人[18] 研究的问题是情绪分类，如图2-1所示，他们通过多维度的模型来构建情绪的数据集，他们的分类维度和 Plutchik 所提出情感轮模型[1] 一致，将所有的情绪分类成八个类别，包括生气 (anger)、害怕 (fear)、厌恶 (disgust)、信任 (trust)、开心 (joy)、惊讶 (surprise)、希望 (anticipation) 和难过 (sad)。Kant 等人所构建的另一个情绪分类数据集[19] 也与此采用相同的分类模型。Lin 和 hsieh[20] 等人研究反讽识别问题，他们通过网络众包的方式来构建数据集，在 Lin 等人的数据集中，他们将反讽的识别任务设计成了一个两类

的文本分类问题，即反讽和非反讽。在 Sem-Eval2013[2] 和 Sem-Eval2017[21] 的情感分析任务重，如图2-2他们都将情感分析 (Sentiment Analysis) 问题设计成了一个三分类问题，分别为正面 (Positive)、负面 (Negative) 和中性 (Neutral)。Ohman[22] 等人构建了态度和情绪的数据集，并且使用 SVM 或者 BERT[9] 对数据集的质量进行评估。

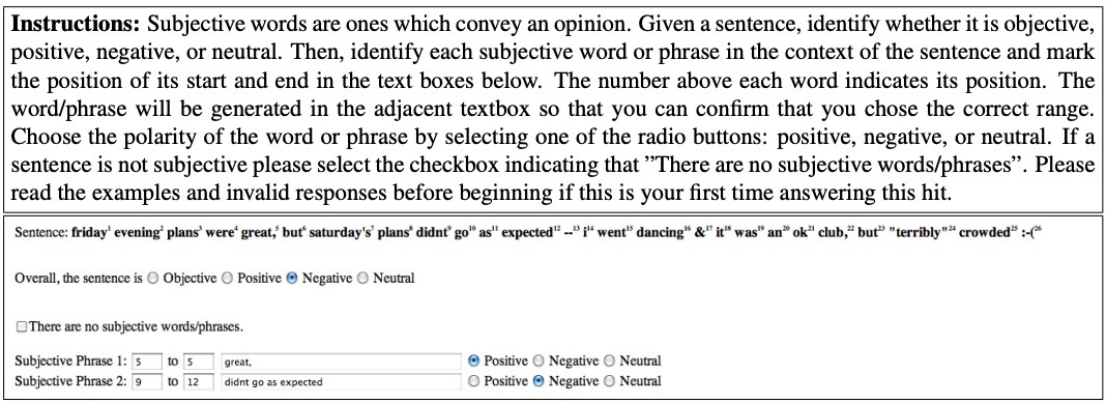


图 2-2: Sem-Eval2013 的标注示例[2]

2.2 句子级别的文本分类

2.2.1 经典机器学习方法

经典的机器学习方法例如支持向量机 (Support Vector Machine, SVM)、逻辑回归 (Logistic Regression, LR) 等线性分类模型，通过划分超平面或是决策边界的形式，对样本进行分类；如朴素贝叶斯 (Naive Bayes, NB) 通过后验概率最大化的方法来确定分类结果等等。在文本分类上相关的模型应用亦十分广泛，通过人工设计每个样本的特征，然后使用相关分类的模型进行分类。Shanahan 等人[23], Mahdi 等人[24], Keerthi 等人[25] 都使用 SVM 作为分类器进行文本分类任务。Su 等人[26]，Dai 等人[27] 使用 NB 作为分类器在设计的特征上进行分类。Wahiba 等人[28] 使用决策树 (Decision Tree, DT) 对文本内容进行分类。Aseervatham 等人[29] 和 P. Jurka Timothy[30] 采用逻辑回归 (Logistic Regression, LR) 对文本内容进行判别。Pranckevicius 等人[31] 将 NB、DT、SVM、LR 和随机森林 (Random Forest, RF) 等方法应用在文本评论分类问题上，并对他们进行详细地比较。经典的机器学习方法通过对特征进行设计，在各自任务上可以有不错的效果。而且 NB、DT、LR 和 RF 等模型还兼具模型推理速度快的特点。

但是，经典的机器学习方法缺点也很明显：首先，特征的设计耗时耗力；其次，模型的泛化性不足。

2.2.2 神经网络方法

情感分析旨在分析用户的观点、情感、价值、态度和情绪等[32,33]。在Bataa[34]、schmitt[35] 和 Tang 等人[36] 的工作中，他们都将情感分析建模成了一个句子级别的分类问题，其中较为粗粒度的问题是，给定一个句子通过模型去判断其是表达的情感是否是正面的；更为细粒度的问题是，给定一个句子，同时给定句子中的方面词 (Aspect words)，去判断每一个方面词对应的描述是否是正面。schmitt[35] 结合了卷积神经网络 (Convolutional Neural Networks, CNN) 和 FastText[37]，构建了一个基于词素卷积的情感分析模型。Tang 等人[36]、Luo 等人[38] 和 Bao 等人[39] 均使用了双向的长短词记忆模型 (Bidirectional Long-Short Term Memory , Bi-LSTM) 和注意力机制 (Attention Mechanism) 作为他们工作中的特征抽取模块。其中 Tang[36] 在 Bi-LSTM 和 Attention 的基础之上增加了额外的监督任务来提升情感分析的效果。Luo[38] 为情感分析设计了与之相关的情感嵌入 (Sentiment Embedding) 和对应的语义嵌入特征 (Semantic Embedding) 来增强情感相关的信号。Bao 等人[39] 修改了注意力机制的产生的损失 (Loss)，给注意力机制的参数添加了不同的正则化方法 (Regularization)。为了可以更好的融合上下文信息 (Context Information)，Liang 等人[40] 引入了上下文相关 (Context-aware) 的嵌入信息，更好地融合了上下文信息的特征和目标序列的特征。

反讽分析 (Sarcasm detection)[41–43] 可以视作是情感分析的一个子领域. Zhang 等人[44] 使用 GloVe[45] 作为预训练的语言模型，结合循环神经网络 (Recurrent Neural Networks, RNNs) 作为特征抽取模块对 tweet 上的文本进行分析。Tay 等人[46] 提出将序列的特征和内部的交互特征分别建模，将序列的特征称为 Sequential Feature, 内部的特征交互称为 Intra-attention，并且用实验证明了，当文本序列经过 RNNs 类似的模型建模之后，再输入到类似于 Attention 的模块中，它们的输出将会变得非常类似，而本文也对此做了相关的实验佐证了这一点，具体内容请见下文。Hazarika 等人[47] 使用了 CNNs 作为特征抽取器，并且结合了最大池化 (Max-pooling) 技术分别建模了文本的内容特征和用户的特征，包括上下文特征 (Contextual Feature)、内容特征 (Content-based Feature)、风格特征 (Stylometric Feature) 和对话特征 (Discourse Feature)，在反讽分析上取得

了长足的进展。

比喻分析 (Metaphorical analysis) 聚焦于探索两个不同概念或者两个不同领域词汇之间的关系[48]。比如“他动也不动，彷如石像。”这句话中，“石像”是一个喻体，本体是“他”，而比喻分析的目的是为了能够找出这一对词，并且判定它们之间的关系是喻体与本体之间的关系。比喻判断（检测）(Metaphor detection)[49,50] 是比喻分析的重要环节, 它的任务目标就是判断给定的文本中是否存在比喻的修辞手法。更为细粒度的比喻分析任务[51,52] 则是去识别给定文本中，哪一个部分是比喻的喻体和本体，这个任务也称为词级别的比喻分析任务 (Token-Level Metaphor Analysis)。Mao 等人[53] 利用 WordNet[①]抽取了语义特征，并且计算了词与词之间的相似度，通过相似程度来判断句子中是否存在比喻。Mao 等人[54] 对文学作品进行了分析，预训练文学特征嵌入 (literal embedding)，然后使用 Bi-LSTM 和 Attention 作为网络的主要模块分析是否具有比喻。

2.2.3 句子级别的多任务建模方法

Majumder 等人[55] 在情感分析任务中引入了多任务的架构，在其结构中，使用双向的门循环单元 (Bi-directional Gated Recurrent Unit, Bi-GRU) 和 Attention 作为特征抽取器，利用一个公共的张量网络 (Tensor Network) 对不同任务之间的特征进行融合。Jin 等人[56] 针对不同的商品类型及其评论进行分析，得出用户评论中蕴含的好/差评观点，因此该文中的多任务架构是为了同时判别分析对象所属的商品类型并且分析其表达出来的情感。Jin 等人[56] 使用 LSTMs 抽取全局的特征，使用多维度的 CNNs(Multi-scale Convolutional Neural Networks) 抽取局部特征，将二者结合作为最后的特征。为了能够记录不同数据之间的关系，该文使用一个全局的张量网络，以使得网络在训练时可以记录所有的数据特征。Akhtar 等人[57] 同时训练两个任务：词抽取 (Term-Extraction) 任务和情感分类任务，相较于单任务的方法有所提升。

2.3 片段级别的文本分类

片段级别的文本分类问题可以直接通过对片段的特征进行建模直接用分类模型进行分类，也可以通过后处理的形式从词级别入手解决。例如，识别一个实体，可以通过识别实体的开头、中间、结尾，然后将开头、中间和结尾进行

[①]普林斯顿大学开源的词林库，其内构建了词的层级关系，包括上下义词等

拼接得到一个完整的实体，从而实现对片段级别的文本进行准确识别。具体内容在本节后续内容中进行描述。

2.3.1 平展类型的片段识别方法

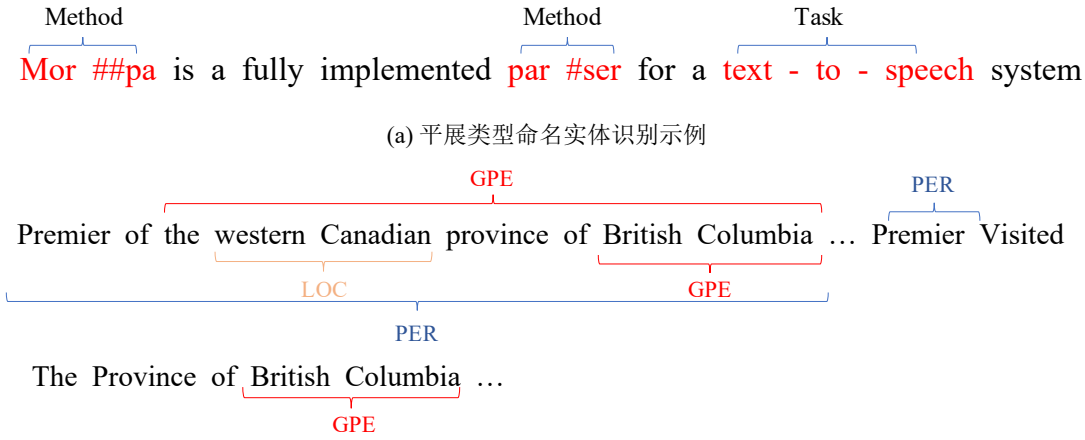
平展类型 (FLAT) 的序列标注是一个非常经典的片段识别方法，其中代表性的任务为 FLAT 类型的命名实体识别 (Named Entity Recognition, NER)，以下简称为 FLAT-NER。它的目标为输入一段文字，输出这段文字中所蕴含的所有实体，根据领域的不同，实体的类型也不相同。Yan 等人[58] 分析了 Vanilla Transformer[59] 在 NER 任务上表现不好的原因，改进了其中关于位置编码 (Positional Embedding) 的设计，采用相对位置编码 (Relative positional encoding) 和无缩放因子的注意力机制 (Un-scaled Attention) 的方式，实现了在 FLAT NER 上的提升。Xuan 等人[60] 提出的 ERINE 使用多种粒度的掩码模型 (Masking Model) 实现知识集成，其中包括关于实体级别 (Entity-Level) 的掩码，在 NER 的任务上取得了不错的效果。Ma 等人[61] 提出使用 ExSoftword 方法来聚合词素信息 (Lexicon)，解决了中文 NER 任务中分词错误的问题和使用 Lattice LSTM 的计算复杂不利于实时应用的问题。Li 等人[62] 通过将 Lattice LSTM 扁平化的方式来解决 lattice 的实时性问题，并且使用相对位置编码避免了 Vanilla Transformer 的问题[59]。

2.3.2 嵌套类型的片段识别方法

相对于 FLAT 类型的片段识别，另一个大家广泛关注的任务为嵌套结构的片段识别。代表性的任务为嵌套的命名实体识别 (Nested Named Entity Recognition, Nested NER)。Nested NER 和 FLAT NER 的区别如图 2-3(a)和图 2-3(b)所示，在 FLAT NER 之中，实体与实体之间是不相交的；而 Nested NER 的任务所要分析的目标，实体可能存在不相交、相交、嵌套等多种情况。

2.3.2.1 基于边界预测的方法

Zheng 等人[63] 针对 Nested NER 任务，提出基于边界预测的方法，该方法的第一步预测候选实体中每一个词素 (Token) 的所属标签，这些标签分为“B”、“I”、“E”、“O”四种，分别代表了一个实体的开头 (Begin)、中间 (Inside)、结尾 (End) 和非实体 (Outside) 四种类型。第二步是组合预测的所



(a) 平展类型命名实体识别示例

(b) 嵌套类型的命名实体识别示例

图 2-3: 两种类型的命名实体识别示例

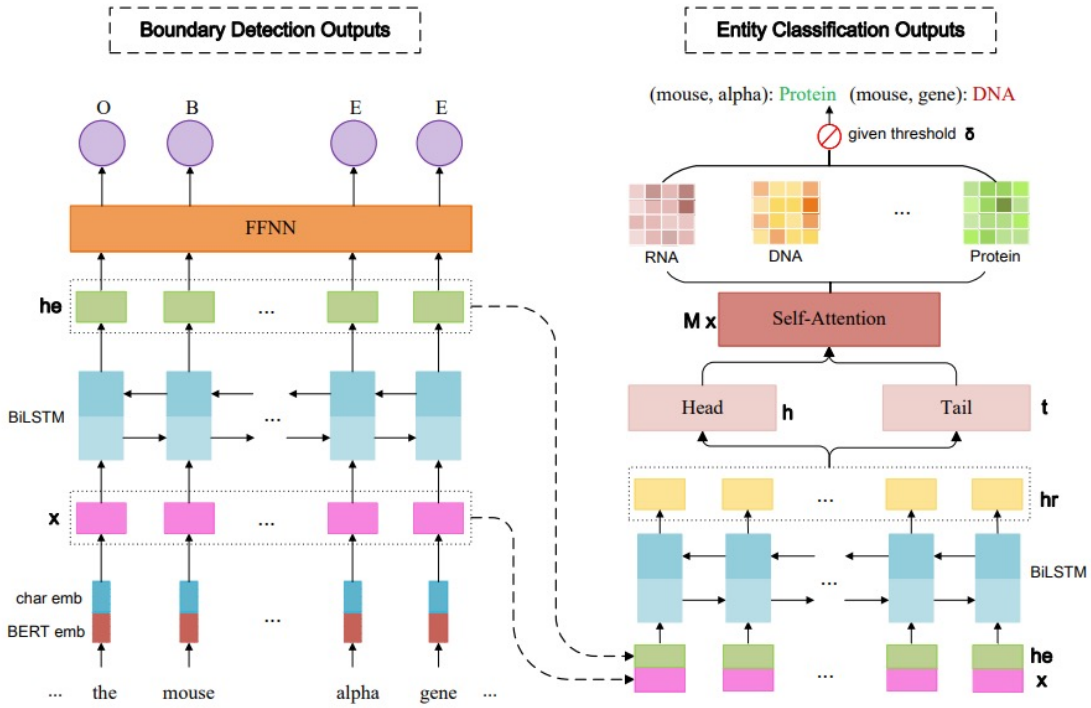


图 2-4: Xu *et al.* 提出的模型结构[3]

有“B”标签和“E”，构成一对标签 (B, E)，以此来代表一个可能的候选实体 (Span)，将候选实体传递给下游进行实体的类型判断，为了避免错误的实体进入候选集，实体分类时通常会加上空实体这一类别。Tan 等人[64] 也采用了类似的方法进行解决，与 Zheng 等人[63] 不同的是，Tan 等人不是去预测一个 Token 所可能的标签类型，而是直接采用两个二分类器，一个预测 Token 是否是一个

实体的起点，一个预测 Token 是否是一个实体的终点，然后对它们进行组合。Xu 等人[3] 对于边界的建模与 Zheng 等人[63] 和 Tan 等人[64] 略有不同。Xu 等人[3] 预测了每个 Token 的标签类型，却不是直接用于实体构建与分类，而是取其对应的特征向量传递到标签分类模块。其模型结构如图2-4所示。Xu 等人[3] 将所有的 Token 映射到了起点空间和终点空间中，然后对其进行组合，然后将组合后的 Span 经过 Self-Attention 的结构映射到不同的实体类别空间中进行分类。将所有类别的预测结果与阈值相比，超过阈值的则取概率最大的类别为 Span 对应的实体类别。Lin 等人[65] 从另外一个角度进行建模，他们认为所有的实体都应该有一个关键的 Token，称之为 Anchor。Anchor 可以代表一个实体的中心位置，Lin 等人[65] 根据 Anchor 去扩展边界。这类给予边界预测的方法非常直观且容易理解，但是其难以避免错误传播问题 (Error Propagation)，即当候选实体的起点和终点这个任务学习不好时，其会直接影响下游的实体判别任务。Xu 等人[3] 虽然没有直接传递标签预测的结果，从而避免了 Error Propagation 带来的影响，但是其在对起点和终点进行组合时需要消耗大量的资源，且多任务的架构更加剧了这种资源的消耗。

2.3.2.2 基于层级结构的方法

层级结构的主要思想是每一层代表一种长度或者一种长度位置的实体，它的代表性结构如图2-5所示。Ju 等人[66] 针对 Nested NER 任务所提出基于堆叠 FLAT NER 模型的方法与图2-5类似。该方法从最小粒度即 Token，开始分析，通过 FLAT NER 的模型，识别这种粒度下的序列标签，得到“B”、“I”、“E”、“O”构成的标签序列。基于识别的序列标签，可以首先获得一部分的实体，即从“B”到“E”为一个实体。将这部分实体采用平均池化 (Mean Pooling) 的方式，得到一个抽象的 Token 表征 (Representation)。然后将该抽象的 Token 替代原来实体所在的位置，得到新的 Token 序列。使用新的一层 FLAT NER 的模型进行重复操作。Wang 等人[4] 的做法，即图2-5所示的方法，在每一层不对实体进行筛选，而是认为在第 i 层，那么这一层所需要判断的就是长度为 i 的实体。与此同时，为了更好的融合多种长度的 Span 信息和避免高层模型产生的梯度无法回传，使用了从高层到低层 LSTM 模型对所有长度候选实体进行了全局建模。Tan 等人[67] 将 Nested NER 建模成成分树的形式，与 Wang 等人[4] 类似，通过树状结构中的节点去代表不同的 Span。除此之外，Tan 等人[67] 提出使用双线性仿射函数 (Biaffine function) 的方式去建模连续

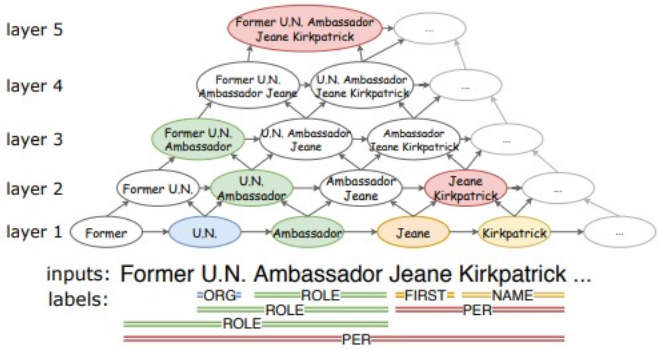


图 2-5: Wang *et al.* 提出的层级模型结构[4]

位置 i 到 j 是一个实体的概率。同时，Tan 等人[67] 的文中将所有的实体分成三种类型：可兼容实体、不可兼容实体和可观测实体。可观测实体即已经被标注出来的实体；可兼容实体是没有在数据中标注出来，但是与标注出来的实体不冲突的候选实体；不可兼容实体是未被标注，但是却与已经标注实体相冲突的实体，这部分实体可以去除以减轻模型的训练压力。

2.3.2.3 基于阅读理解的方法

阅读理解的主要形式是给定一个问题，从文章中找出一个答案。基于阅读理解的方式和阅读理解的形式是一样的，这种方法先行构建了标注相关的问题描述，然后将该问题与原文段进行拼接，中间用一个特殊的符号进行区分。然后从文段中识别出相应的内容。Yan 等人[68] 在他们的文章中，将所有的问题都设计成了问句的形式，例如要抽取“PER”(人名，Person) 实体，问句则为：“Which person is mentioned in the text?”。当存在两个嵌套或者相互交叠的实体时，它们会被不同的问句给区分开，因而避免了嵌套识别的问题。然而，这种方式也有一定的局限性，如果问句的构成是固定的，即每次只是替换了问句中所蕴含的类别信息，当两个交叠的实体或者嵌套的实体属于同一个类型时，在阅读理解的问句设定下，依旧还是一个嵌套实体的问题，依旧没办法解决。如果这种问句是特殊设计的，当遇到两个交叠的实体是同一类型，但是却是不同问题的答案时，可以被解决。然而这种问句的设计是非常消耗人力的，它需要针对所有可能的情况去定制化一系列的问句。

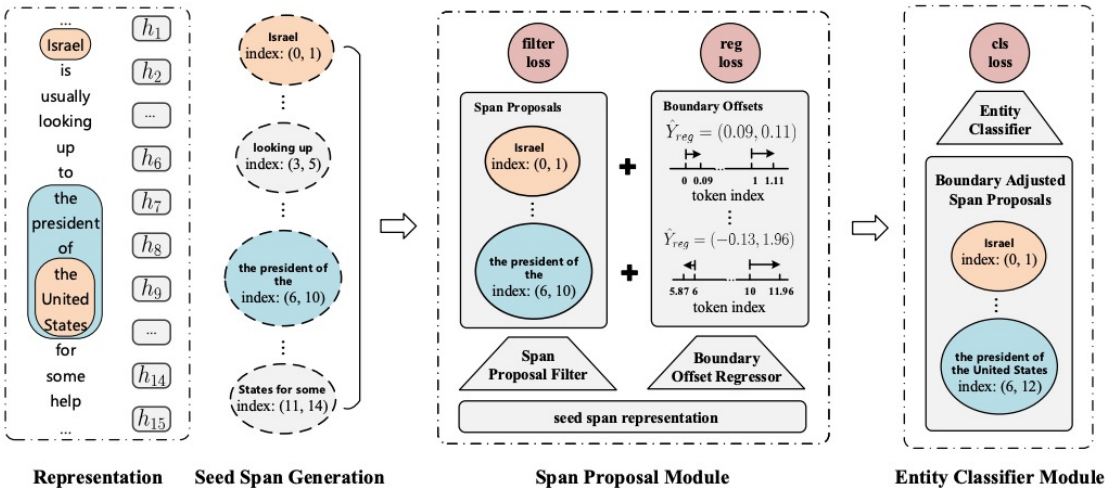


图 2-6: 两段式方法的模型结构[5]

2.3.2.4 基于两段式的方法

为了更好的描述两段式的方法，此处先引入一个新指标——Intersection-over-Union (IoU)。IoU 描述的是两个 Span 之间的交叠关系。给定两个 Span $s_1 = \{w_1, w_2, \cdots, w_m\}$ 和 $s_2 = \{x_1, x_2, \cdots, x_n\}, w_i, x_i \in \mathbf{V}$，其中 $\mathbf{V}$ 是任务的词表。则 Span s_1 和 s_2 的 IoU 值定义为：

$$\text{IoU}(s_1, s_2) = \frac{s_1 \cap s_2}{s_1 \cup s_2}. \tag{2-1}$$

如图2-6所示，Shen 等人[5] 认为嵌套的命名实体识别可以拆分成两个子任务：Span 的过滤和 Span 的分类。其中 Span 的过滤是指第一阶段将一段文本中，所有可能的 Span 进行枚举。从大量的枚举结果中，训练一个二分类器，这个二分类器旨在将所有的 Span 区分成两种：高质量的 Span，称之为 Proposals；低质量的 Span，需要被过滤掉。而高质量和低质量的区分，通过 IoU 进行选择。当一个 Span 和 Entity 之间的 IoU 值高于阈值 α 时，则将这类 Span 定为 Proposal。第二阶段，对 Proposal 进行范围的调整和分类。

2.4 本章小结

本章将本文相关的参考工作分成了三个部分进行叙述，第一个部分为数据集，该节介绍了和数据集相关的标注工作以及标注样式。第二个部分为句子级别的文本分类工作，该节介绍了句子级别的情感分析、反讽识别、比喻分析和

一些多任务的模型。第三个部分为片段级别的文本分类工作，该节将片段级别的文本分类模型分成两种类型进行介绍：FLAT 类型和 Nested 类型。这三个部分的参考工作分别对应了后文所述的三个章节。

第三章　中文潜台词数据集的构建

本章的内容旨在对本文涉及的数据进行详细地介绍，包括数据来源、搜集、标注、数据质量评估和分析。经过大量地观察和分析，本文发现潜台词(Subtext) 常出现在两种场景：文学作品以及热点媒体事件中的评论。文学作品中的潜台词使用非常广泛，同时潜台词牵扯的内容跨度非常长，常在某个章节中埋下某个伏笔，几个章节之后才能了解其深层次的意思。限于机器的计算性能与模型所能有效处理的长度，本文不对这部分数据进行分析。热点媒体事件中的评论由于其媒介的特殊性，通常长度都不会很长。例如关于某个流行歌曲的评论，其长度通常都在几十个词左右。出于这种长度的限制，其所需要表达的潜台词，很容易的可以从所评论的主题或者相关评论之中获取。因此，本文的研究内容主要针对该种非正式化的文本内容。

3.1　数据搜集

本文的数据都搜集自微博、知乎、网易云音乐和哔哩哔哩的热门榜单，这些榜单具有较高的关注度和较多的评论信息。为了可以让数据更具有泛化性和可用于更多的研究分析，本文保留了源数据的一些结构信息，包括当前评论信息 (Comment information)、当前的评论 ID(Comment ID)、父评论信息 (Parent Comment)、父评论 ID(Parent ID)、来源信息 (Source Information)、主题信息 (Theme Information)，其中父评论信息定义成如果一条评论被另一条评论所评论，则该评论的信息为父评论信息，评论该条信息的评论可称为子评论信息。本文中保留这种信息的目的有两个：1. 期望可以给评论信息以更丰富和完整的上下文；2. 期望研究者可以通过这种关联，对多轮 (Multi-rounds) 条件下的潜台词进行研究。最终，本文从这些网站中搜集到 72,494 条评论数据。

3.2　标注

表 3-1: 原始数据

No.	comm	p-comm
1	T+0 券商手续费赚到手软。。。。	众所周知，A 股有 2 条独特的交易制度：⋯ 是何居心了吧？
2	学校不就是一个大型内卷培养皿吗？	外卖行业对准时性的要求越来越高，⋯ 做不到的地步了。
3	貌似现在收入 ⋯ 还是大部分，但是房价已经翻了好几番了 [惊喜]	他可能薪资优渥,⋯, 住房的需求仍然会不断增加。
4	如今羊村群神争霸，我喜羊羊是个一无是处的废物？⋯，还是要由我白嫖党来拯救！！	NULL
5	是五常给联合国权利，不是联合国给五常权利 [吃瓜]	NULL

表 3-2: 数据集标注示例

No.	subt	sarc	meta	exag	homo	emot	sent
1	1; 赚到手软。。。。;T+0 对用户更加不好	-1;-	-1;-;-	1; 赚到手软	-1;-;-	0	1
2	1; 大型内卷培养皿; 学校的竞争就很激烈	-1;-	1; 大型内卷培养皿; 培养大量类似人员的地方	-1;-	-1;-;-	2	0
3	1; 房价已经翻了好几番; 大部分人买不起房	1; 现在收入 ⋯ 还是大部分，但是房价已经翻了好几番了 [惊喜]	-1;-;-	-1;-	-1;-;-	2	0
4	-1;-;-	-1;-	1; 白嫖党; 伸手不花钱族	-1;-	-1;-;-	6	0

5	-1;-;-	1; 不是联合国给五常权利	-1;-;-	-1;-	-1;-;-	0	0

为了保护个人隐私，本文进行标注前，已经对数据进行预处理，将与用户私人相关的信息如用户 ID、用户昵称等去除，仅保留广为人知的公众人物的名字。标注的部分示例如表 3–2所示，其所对应的原始数据可根据 No. 在表 3–1中查看。其中 comm 和 p-comm 分别表示评论和父评论。本文针对每条评论信息，标注了七类信息：潜台词 (subt)、反讽 (sarc)、比喻 (meta)、夸张 (exag)、谐音 (homo)、情绪 (emot)、观点 (sent)。其中，“其它”是一个笼统的标签，包含了除单独列出的几类表现手法以外的其余所有手法。“其它”中所包含的表现手法，在本文的数据集中的数量非常少，经统计不足 50 条，因而将这类表现手法都归拢到一个统一的标签下。潜台词、反讽、比喻、夸张、谐音等都具有类别标签，将该标签放在标注信息的第一位；标注信息的第二位代表具有这种标签倾向的原文内容。对于潜台词、比喻和谐音，还有第三位标注信息，这类信息代表的是它们真实要表达的内容。例如，在表 3–2中，潜台词的第三位标注信息“T+0 对用户更加不好”，代表的是“赚到手软。。。。”这个内容想表达的内在含义；比喻的第三位标注信息“培养大量类似人员的地方”代表的是“大型内卷培养皿”这个喻体所对应的“本体”。每一位标注信息通过 “;” 进行分割。对于标注信息的第一位，本文效仿相关工作中的做法，将它们标注成三类：“-1”代表该条数据不包含“潜台词”、“比喻”、“反讽”、“夸张”和“谐音”;“0”代表该条数据不确定是否有这类信息；“1”代表该条数据明确包含“潜台词”、“比喻”、“反讽”、“夸张”或“谐音”。据调查显示，本文也是首个标注出“夸张”和“谐音”修辞手法的工作。除此之外，本文针对每条数据，标注了评论者评论时的情绪信息和观点态度，因此我们的数据集可以应用在更多的任务类型中。其中情绪信息，我们效仿 Mishra 等人[18]，将它标注成了八个类别外加“无”这个第九类别。我们将这九个类别分别对应到了九个数字中，“0”代表“无”，“1”代表生气 (anger)、“2”代表害怕 (fear)、“3”代表信任 (trust)、“4”代表厌恶 (disgust)，“5”代表难过 (sad)、“6”代表开心 (joy)、“7”代表惊讶 (surprise)

和“8”代表希望 (anticipation)。对于观点态度，我们效仿 Sem-Eval2013[2]，将其标注成三类：“-1”代表反对、“0”代表无明显倾向/客观、“1”代表支持。最终，经过对标注数据去重和去除无效信息，我们得到了 8,843 条标注数据。

3.3 质量评估

为了确保本文标注工作的质量和评估其有效性，本文采用了两段式的标注方法 (Two-Stages Labeling Methods) 来缓解标注时产生的主观性影响和采用了两种方式来评估 CSD 数据集的可靠性。同时考虑标注成本和主观因素的影响，本文采用的两段式标注方法分成以下两个步骤：第一轮标注时，每条数据都经由三个人独立的进行标注；第二轮标注时，由第四个人对第一轮标注的三个结果进行审核。审核分成以下三种情况：

- 如果所有标注结果一致，则审核通过，将标注结果视为黄金标签 (Golden Label) 作为最终的标注的结果，并且对标注信息的第二位和第三位进行完善。
- 如果部分标注结果一致，则由第四位人员根据第一轮标注的结果进行评判，选择重标或者选取其中一个结果作为黄金标签，并且对标注信息的第二位和第三位进行完善。
- 若所有结果不一致，则舍弃该条数据。

为了衡量数据的可靠性，本文效仿 Ghanem 等人[69]、Khodak 等人[70] 和 Webster 等人[71] 使用 Kappa Score 作为衡量的一项指标。然而，Kappa Score 通常是作为单轮完全独立的标注条件下的衡量手段，这个与本文中的标注方式有些出入。更有甚者，Kappa Score 在数据极度不平衡的条件下的结果具有欺骗性[72,73]: 当数据极度不平衡时，尽管标注者有着非常高的一致性，Kappa Score 的数值也可能很低，导致对数据质量得出错误结论。因此，本文提出另一种衡量方式称为两阶段标注衡量方法 (**T**wo-Rounds **A**nnotation **E**valuation, TAE)。并且引入 Ohman 等人[22] 的数据集评估方法，使用 SVM 作为基本分类器，将数据集分成训练集和测试集，使用测试集上的评测结果对数据质量进行衡量。

TAE Score: 为了计算 TAE Score，首先为单条数据单个类型定义两个数值。

- 一致性 (Agreement)：第一轮标注的三个标注结果集和第二轮标注结果集相同的数量占总数量的比值。设 $L_1 = \{l_{11}, l_{12}, \cdots, l_{1n}\}$ 为第一轮标注的结果集，

设 l_2 为第二轮标注的结果，其集合表示为 $L_2 = \{l_2\}$，其中 n 为第一轮标注的人数。则对于第 i 条数据，它的一致性计算如下：

$$agr_i = \frac{|||\{l_{1j}|l_{1j} \in L_2, j = 1, 2, \cdots, n\}|||}{n}. \tag{3-1}$$

- 随机性 (Randomness): 设两轮标注结果的总类型数为 **T**，在第一轮标注结果中出现但未在第二轮标注结果中出现的标注类型数量为 **N**，则随机性定义成 **N** 和 **T** 的比值。设 ls(s) 是将一个列表 s 转换成集合的函数，设 $Li_1 = [l_{11}, l_{12}, \cdots, l_{1n}]$ 是第一轮标注的结果，设 $Li_2 = [l_2]$ 是第二轮标注的结果，则第 i 条数据的随机性计算如下：

$$rad_i = \frac{\mathbf{N}}{\mathbf{T}} \tag{3-2}$$

$$\mathbf{N} = \text{ls}(\text{Li}_1) \backslash \text{ls}(\text{Li}_2) \tag{3-3}$$

$$\mathbf{T} = \text{ls}(\text{Li}_1) \cup \text{ls}(\text{Li}_2). \tag{3-4}$$

作为可靠性衡量的方法，TAE 应该满足以下三条性质：

- 单调性 (Monotony)：TAE Score 关于一致性应该要满足单调递增的性质，同时应该要关于随机性满足单调递减的性质。
- 有界性 (Boundness)：TAE Score 应该关于随机性和一致性有确定的上下界，只有拥有确定的上下界，我们才能衡量一个数据集是否是好的。
- 独立性 (Independence)：TAE Score 的数值应该跟标注数据的标签分布是相互独立的，即不应该随着正负样本比的变化而变化，而这个是 Kappa Score 在上文中所提到的主要缺陷。

因此，根据上述性质，本文将 TAE 定义如下：

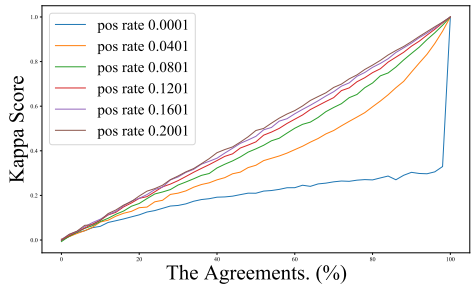
$$\text{TAE} = \frac{\exp(agr - rad) - 1/e}{e - 1/e} \tag{3-5}$$

$$\text{agr} = \frac{\sum_{i=1}^{n} agr_i}{n} \tag{3-6}$$

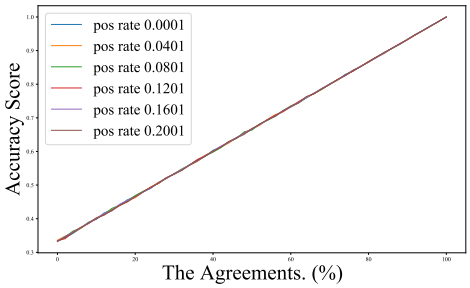
$$\text{rad} = \frac{\sum_{i=1}^{n} rad_i}{n}, \tag{3-7}$$

其中 e 为自然底数。显然，TAE 满足单调性，又由于 $agr, rad \in [0, 1]$，可知 $TAE \in [0, 1]$，因此 TAE 也满足有界性。TAE 的独立性本文后文通过模拟实验进行验证，如图3-1所示。

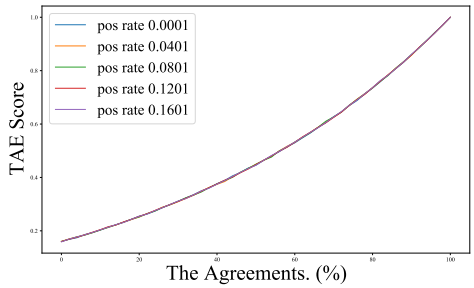
为了衡量 TAE 的有效性，本文做了相关的模拟实验。本文在不同的正负样本比的条件下，进行了六组实验，并将 Kappa score、正确率 (Accuracy) 作为对比。每组实验进行的任务为三分类任务，这与本文所进行的任务相契合。实验结果如图 3-1所示。 图 3-1(a)、图 3-1(b)、图 3-1(c)展示的是 Kappa score、



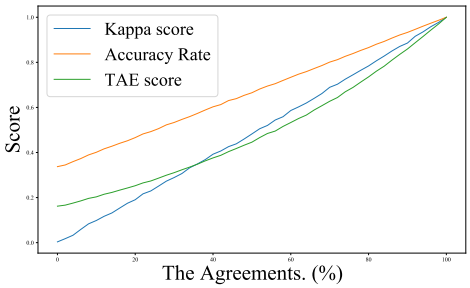
(a) Kappa Score 随着一致性变化的图示



(b) Accuracy Score 随着一致性变化的图示



(c) TAE Score 随着一致性变化的图示



(d) 相同正负样本比下的不同衡量方法变化图示

图 3-1: 不同衡量方法比较结果展示

Accuracy score 和 TAE score 的曲线随着一致性的变化情况，图上的不同曲线，代表着它们是在不同正负样本比的条件下所做的实验。图 3-1(d)展示的是在正负样本比为 1：5 的条件下，Kappa score、Accuracy score 和 TAE score 它们的变化情况。其中图 3-1(b)表明正确率的曲线随着一致性的增加而呈现线性变化，并没有考虑到随机性带来的影响。而图 3-1(a)和图 3-1(c)表明 Kappa score 和 TAE score 关于一致性呈现非线性变化，二者都考虑了随机性在评估中起的作用。从图 3-1(a)中，可以发现 Kappa score 的分数曲线不是固定的，而是会随着正负样本比的变化而发生改变，尤其当正负样本比很小时，即数据不平衡时，它所呈现出的 Kappa score 当 Agreement 分数很高时，也会将数据集判定成不可靠，而这显然不合理。而从图 3-1(c)中，可以发现，TAE score 几乎不会随正负样本比的变化而变化，对于不同的正负样本比它都能够呈现类似的结果，即满足上文所述的独立性。且从图 3-1(d)中可得，在数据较为平衡时，TAE score

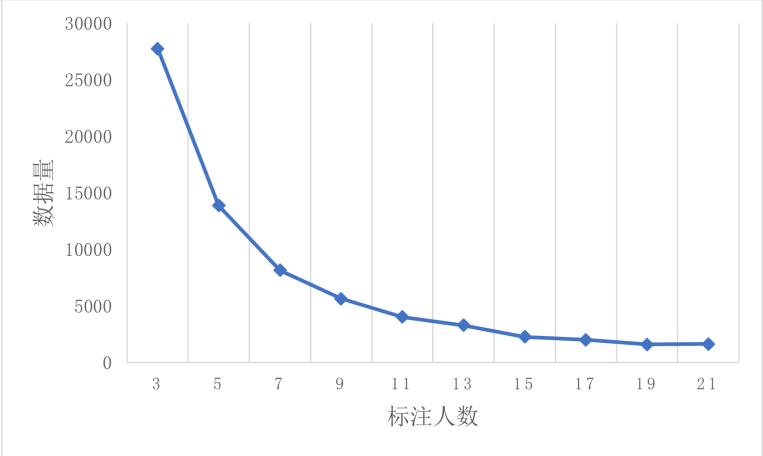


图 3-2: 每条样本的标注人数统计

和 Kappa score 的衡量效果相当。且图 3-1(d)表明正确率在相同正负样本比条件下比 Kappa score 和 TAE score 更高。当一致性低于 0.35 时，TAE score 的数值要高于 Kappa score，而后要比其更低。总体而言，TAE 有接近 Kappa score 的效果。因此，TAE 可以用于辅助判别我们的数据集是否可靠。同时，我们通过二分法近似得出，当 Kappa score 的数值在 0.6 时，所对应的 TAE score 大约在 0.53，这意味着，Kappa score 中一个数据集如果高于 0.6 是可靠的，那么在 TAE 中，当分数高于 0.53 时，可认为其可靠。

为了进一步验证，本文所提 TAE score 的合理性和有效性，本文在 Amazon 提供的公开数据集[①]上使用 Kappa score 和 TAE score 进行测试。由于 Amazon 的数据是通过用户自动化进行评分得到，而 TAE score 在评分时主要针对两轮标注进行衡量。而 Amazon 的数据相当于仅有一轮数据，因此，在对其进行衡量时，需要根据对每个评论的所有评分进行投票，然后得到一个最终评分，将该最终评分作为第二轮标注的结果。虽然其是一个打分的标注，然而每个分都是一个确定的整数，范围在 1-5 之间。因此，可以将其看成是五分类问题。如图 3-2所示，每个样本的标注人数随着人数的增加而急剧减小。同时考虑样本类别与标注人数的影响和与本文所研究问题的相似性，对 Amazon 的数据作出如下处理，将 3 分以下的数据类别定义为差 (-1)，3 分的样本定义为一般 (0)，3 分以上样本定义为好 (1)。

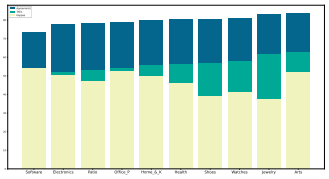
基于上述处理后，本文对 Amazon 的数据可靠性进行评估，衡量了不同标注人数的数据中，不同子领域之间的数据可靠性和总体数据集的可靠性。衡

[①]数据集地址：http://snap.stanford.edu/data/web-Amazon-links.html

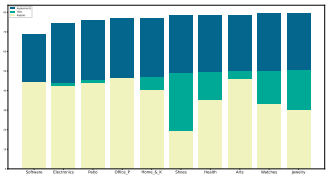
表 3-3: Amazon 数据集的衡量结果

标注人数	子数据名	Kappa score	TAE Score	agreements	1:0:-1
3	Jewelry	37.21	61.34	83.16	1228:49:67
	Arts	51.75	62.55	83.79	289:9:33
	Shoes	38.88	56.47	80.64	3791:157:293
	Home_&_Kitchen	49.69	55.69	79.77	5332:171:809
	Watches	41.29	57.69	80.93	781:19:72
	Software	53.85	45.71	73.21	700:32:317
	Office_Products	52.50	54.30	78.96	976:25:192
	Patio	46.86	52.81	78.30	1443:43:248
	Health	45.90	56.42	80.32	2964:106:373
	Electronics	50.37	52.05	77.55	5820:222:1189
	Average	48.57	54.69	79.27	-
5	Jewelry	30.09	50.46	79.64	578:11:25
	Arts	45.67	49.87	78.69	133:3:17
	Shoes	19.22	48.90	78.57	2035:46:52
	Home_&_Kitchen	40.32	46.79	76.75	2718:79:354
	Watches	32.77	49.94	79.51	343:7:19
	Software	44.03	37.42	68.98	392:13:143
	Office_Products	46.45	46.37	76.69	486:10:84
	Patio	43.80	44.98	76.04	742:25:116
	Health	35.00	49.14	78.66	1534:23:131
	Electronics	42.33	43.77	74.48	3134:85:555
	Average	39.40	46.29	76.51	-
9	Jewelry	11.63	38.81	74.81	143:0:3
	Arts	46.16	48.41	83.20	37:0:4
	Shoes	22.18	42.27	81.89	797:1:20
	Home_&_Kitchen	35.03	38.91	74.59	1242:5:151
	Watches	15.91	45.86	80.25	123:0:3
	Software	43.79	32.99	68.38	159:1:67
	Office_Products	41.35	38.08	73.94	231:2:42
	Patio	32.89	38.62	74.14	286:2:36
	Health	30.18	40.49	76.13	644:5:53
	Electronics	37.37	37.99	73.34	1356:9:235
	Average	35.86	39.21	75.37	-

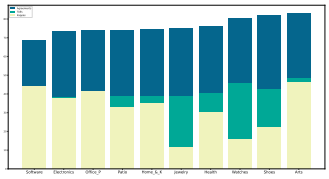
量结果如表3-3所示，其中最后一列为类别分属 1、0 和-1 的数据量之比。为了更好地展示衡量的结果，我们对表3-3的数据，按照一致性从低到高进行排列，将一致性 (蓝色部分)、TAE score(绿色部分) 和 Kappa score(黄色部分) 共同呈现在图3-3中。从表中数据可以看出，Amazon 的数据在各个子数据集上的一致性是很高的，总体数据集的一致性也是很高的。且在相同的一致性条件



(a) 3 人标注的衡量结果可视化



(b) 5 人标注的衡量结果可视化



(c) 9 人标注的衡量结果可视化

图 3-3: 衡量结果的可视化

表 3-4: 不同衡量方法的衡量结果

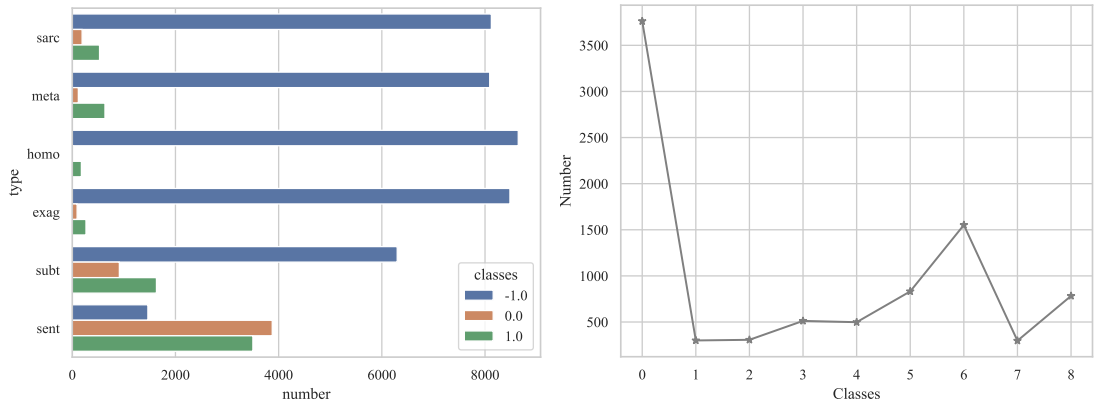
type	sarc	meta	subt	exag	homo	sent
Kappa	0.6028	0.6021	0.6042	0.7065	0.6113	0.7307
TAE	0.8071	0.5618	0.5016	0.8818	0.9515	0.5063
$\mathbf{F}_{1,SVM}$	0.5087	0.4976	0.4657	0.5030	0.5322	0.5421

下，Kappa score 和 TAE score 表现各有不同，当数据标签不平衡的时候，Kappa score 对数据的衡量即存在偏差，容易判成不可靠。而 TAE score 则不会受数据标签的分布是否平衡的影响，依旧可以比 Kappa score 更能准确衡量数据集的可靠性。在数据较为平衡的时候，Kappa score 和 TAE score 都可以得出相似的可靠性结论。从图 3-3中可以更直观的看出，TAE score 和一致性具有更一致的变化关系，而 Kappa score 和一致性的关系较为混乱，并没有呈现出特定的规律。由此可见，TAE score 相比于 Kappa score，更合理的考虑了标注的一致性了，同时从 TAE score 和一致性之间的差别也可看出，TAE score 还考虑了标注的随机性。

3.4 数据分析

经由上文所述，本章节对数据分析的结果进行简要的展示。CSD-数据集关于 Kappa score、TAE score 和 SVM 预测的衡量结果如表 3-4所示。

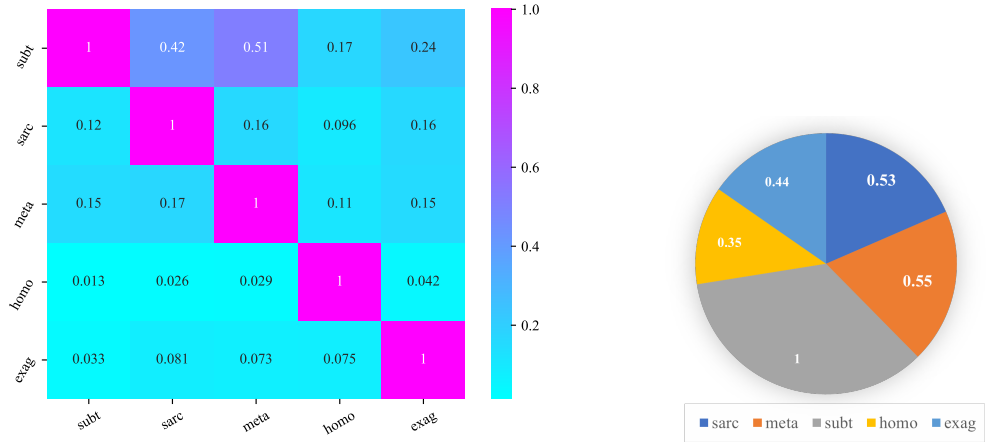
综合三者的分数，可以直观得到 CSD-数据集在反讽、比喻、潜台词、夸张、谐音和态度的有效性，且可得在三个指标的衡量中潜台词的可靠性均在足够可靠的界限边缘，可见潜台词任务的标注任务的难度，即使对人而言，也不是轻易可以准确识别。图 3-4展示了 CSD-数据集的标签统计信息，其中图 3-4(a)展示的是所有三分类标签的类别分布情况。图 3-4(b)展示的是不同情绪的样本分布，可以发现，大部分的样本都没有表达出确定的情绪，表现出的



(a) 三类别标签的分布 (b) 情绪标签的分布

图 3-4: 标签分布

绝大部分情绪为 5(难过)，6(开心) 和 8(希望)，这个体现出来的情绪聚焦特征也非常符合我们的内容来源。



(a) 属于类型 2(横轴) 的所有样本中又属于类型 1(纵轴) 的样本占比 (b) Spearman 系数

图 3-5: 相关性分析展示

另外，我们还对不同标签之间的相关性进行了分析，分析结果如图 3-5所示。其中图 3-5(a)展示的是，在类型为 type2 的所有样本中，包含 type1 的样本比例。其中横轴代表 type2，纵轴代表 type1。可见，在各种样本中，潜台词所占比例都是最高的，其中反讽类型和比喻类型的样本中有将近一半的样本都表示了潜台词，而在潜台词中反讽类型和比喻类型所占比例也是最高的。图 3-5(b)表示的是潜台词和其余各个类型之间的 spearman 系数，发现反讽和比喻与潜台词有较高的相关度，均超过了 0.5，这为下文我们所采用的方法提供

了一些统计的数学基础。

3.5 本章小结

本章对 CSD-数据集进行了详细介绍，包括数据类型的选取、数据的来源与收集方法、数据的标注方法、数据质量的评估方法和数据集地分析。在数据质量评估一节，本文提出了 TAE score，明晰了 TAE score 的定义和其应该满足的性质。本文在 Amazon 数据集上对 TAE score 和 Kappa score 的表现做了相关对比，得出 TAE score 是可靠的，可以用于评判数据是否可靠。最后，本文将 TAE score 用作辅助判断数据是否可靠的一种方式，并且和 Kappa score、F_1 score 共同验证了 CSD 数据集的可靠性。

第四章　句子级分类任务：潜台词检测

本章的内容旨在对潜台词检测以及本文所提出的解决方案进行详细地介绍。包括任务的辨析、模型的设计思路、实验结果及其分析。

4.1　潜台词检测与相关任务的联系与区别

本节的主要内容旨在辨析潜台词检测、情感分析、比喻分析和反讽分析之间的区别。

潜台词与反讽、比喻任务和情感分析有着一定的相似性，从图3-5(a)中可得，它们相互之间具有重叠，但却不同。潜台词是一种语言现象，是一种表达意图的手段，它有着不同的载体，比如通过反讽的手法来进行批判、通过比喻来更形象地表达自己的意图，但是仅有载体不能够说明是否具有潜台词，还要更进一步看能否通过这些载体表现出与表面文字不一样的意图。更进一步，潜台词通过反讽和比喻当作载体，仅仅是其中的一小部分，更多的潜台词是通过当前的上下文关系或者根据一些背景知识即兴而发的，这一部分的潜台词需要我们对语句所处的上下文有更全面和更深刻的认识。

为了更好地说明潜台词与反讽、比喻和情感分析之间的关系，本节将列举一些数据集中的例子进行说明。例如在一首情歌的下面，有这么一句评论“你的心有一堵墙，我要跨过这道墙”，这句话的潜台词很明显——我喜欢你，我想追求你。对于比喻分析而言，它的目标是识别这句话中具有比喻，其中“墙”比喻困难，“跨过墙”比喻克服困难。对于反讽识别而言，它的目标是判断这句话不具有反讽意味。对于情感分析而言，它需要判断这句话的情绪是“希望”，代表的态度或者观点是“中性”即无明显倾向。但是对于潜台词分析而言，它需要判断出这句话具有潜台词。根据“跨过墙”意味着“克服苦难”，而这个“困难”在“心里”，同时这是“关于一首情歌的评论”进行推理，最终得出“我喜欢你”这样具体的意思。例如关于哔哩哔哩某视频下的

评论“1 个月一期，一共一百期，追个几年没问题了 [doge]”的评论为“好家伙，起码 9 年”，这句话反讽分析可得具有讽刺的含义，细化一点可以定位到“好家伙”和“九年”具有很强的讽刺意味。情感分析可以分析出他是**支持**所评论的评论的，情绪是“惊讶”。比喻分析可得这句话没有比喻的成分。对于潜台词分析需要得出，这句话具有潜台词，根据反讽的结果，结合上下文“1 个月一期，一共一百期”，进行推理，可得这句话反映的真实意思为“这个系列的视频更新非常缓慢”。

但不是所有潜台词，都是有反讽或者比喻蕴含其中。例如面对“拼图看到了帅小伙在哪呢 [doge]”这句评论，有人回复说“在这呢 [害羞]”, 这句话在比喻分析和反讽分析中的结果是无比喻也无反讽，但是从上文中“帅小伙在哪呢”这一句可得，该句需要询问的是“帅小伙”，而回复“在这呢”，完整应该是“帅小伙在这呢”，进一步可得答者的意图为“我就是帅小伙儿”而不是“我在这”。该例子中，答者刻意强调了话者的字面意思，对其所要表达的内容“视频上的人不帅”曲解，以达到使用调侃的方式表达“我很帅”的目的。

由本节前文所述可知，比喻、反讽、情感分析虽然从表面意思和潜台词分析有相似之处，但还是有很大的区别存在。前文中的例子，可以很明显的看出，情感分析和潜台词分析差异较大，分析目标也相去甚远，不过均可作为语义理解的一个组成部分。而关于反讽和比喻比较容易让人产生混淆的感觉。对于存在反讽和比喻的句子，有可能是潜台词，且这种时候其都作为了潜台词的语义识别的一个关键部分，但不是全部，只是其中的一个环节而已；更有甚者，潜台词很大一部分都是依靠上下文而产生的，与比喻和反讽并无关联。

本段将反讽、比喻、情感分析和潜台词分析的区别简要概括如下。首先，潜台词检测、情感分析、反讽分析和比喻分析的任务目标不同：情感分析旨在分析目标语句的情绪、观点态度的正负面性；反讽分析旨在分析出当前语句是否具有反讽的意味；比喻分析旨在分析出当前语句是否具有比喻的修辞手法，进一步的任务是找出其喻体与本体；潜台词检测的目的是为了识别句子是否有浮于文字之下的深层次含义，比之前三个任务要更加的困难。其次，潜台词检测比反讽分析、比喻分析和情感分析更加依赖上下文，而且是否有反讽和比喻，仅能作为是否具有潜台词的一个因素，而不是结果。最后，潜台词检测有时候需要有更强的背景知识，因为它在文章中出现的时候可能不会有和比喻分析、反讽分析类似的关键字作为判断的依据。因此，本文重申我们的观点，潜台词分析有着其必要性和前瞻性！

4.2 SASICM 模型结构

本节旨在针对潜台词检测的解决方案做一个详细的介绍。根据章节1.1.1中所述，我们对潜台词检测的建模采取的方案为：学习一个指示函数 $I(x,y)$，当 x 和 y 表示同一个意思时，则指示函数的值 $I(x,y)=-1$；当 x 和 y 不确定是否表示同一个意思时，指示函数的值 $I(x,y)=0$；当 x 和 y 不表示同一个意思时，则指示函数的值 $I(x,y)=1$。同时，根据章节3.4中的数据分析结果显示，反讽和比喻与潜台词有更高的相关度，可以为潜台词提供一些线索，因此本文采用多任务方式来加强潜台词分析的效果。共分成三个任务：识别一句话中是否具有反讽、识别该句中是否具有比喻以及识别该句是否具有潜台词。关于这三个任务如何在模型中进行组织，详见下文。本章将所用模型命名为基于强化注意力机制的混合序列与内部的多任务模型 (Strengthen Attention based Sequence and Intra-Attention Confused Multi-Task Model, SASICM), 其结构如图4-1所示。模型包含了以下几个模块：编码模块 (Encoder)、强化注意力层 (Strengthen Attention)、双向的长短词记忆神经网络 (Bidirectional Long Short-Term Memory)[74]、特征混合模块 (Feature confusion)、语义抽取模块 (Mean Extractor) 和三个任务的分类模块。

4.2.1 问题建模

设 $S=\{<cls>,w_1,w_2,\cdots,w_n,<sep>,c_1,c_2,\cdots,c_m\}$ 为输入的文本序列，其中 $<cls>$ 和 $<sep>$ 是依照 BERT[9] 所设立的两个特殊字符，$W=\{w_1,w_2,\cdots,w_n\}$ 是需要分析的文本内容，$C=\{c_1,c_2,\cdots,c_m\}$ 是当前文本所关连的上下文信息。任务的目标是学习三个函数: 1. 潜台词检测函数：$y_{subt}\leftarrow I_{subt}(W,S)$，$y_{subt}\in\{-1,0,1\}$，分别代表确定没有潜台词、不确定是否有潜台词和确定具有潜台词；2. 反讽判断函数：$y_{sarc}\leftarrow I_{sarc}(W)$，$y_{sarc}\in\{-1,0,1\}$，分别代表确定没有反讽、不确定是否有反讽和确定具有反讽；3. 比喻判断函数：$y_{meta}\leftarrow I_{meta}(W)$，$y_{meta}\in\{-1,0,1\}$，分别代表确定没有比喻、不确定是否有比喻和确定具有比喻。本章设计三个问题同时解决的方案有以下原因：1. 从章节3-5中可以发现，如果一个句子具有反讽或者比喻，那么该句有比较大的可能具有潜台词；2. 期望反讽和比喻可以给潜台词的判断提供更多的线索。

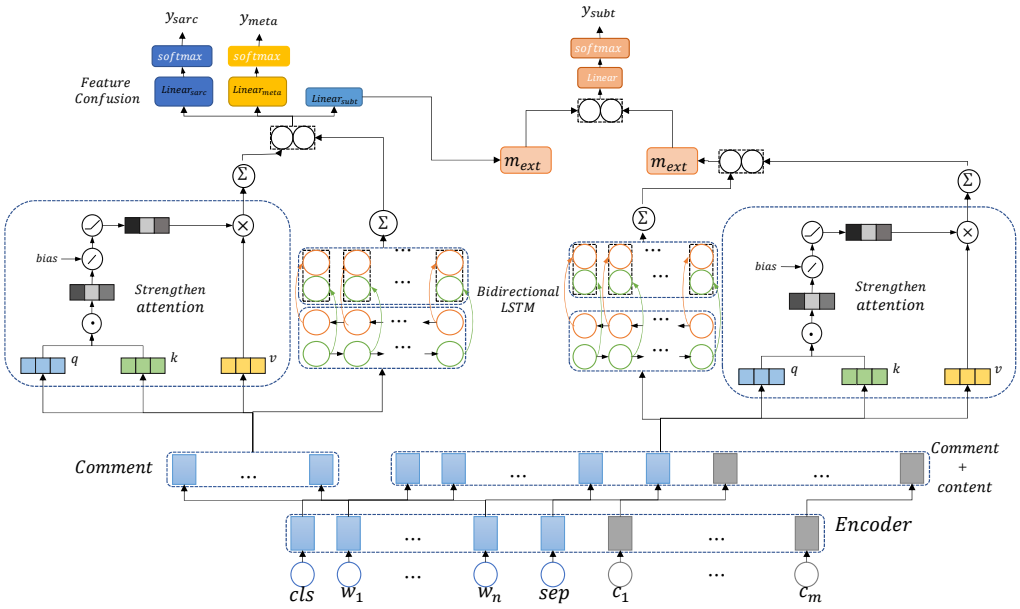


图 4-1: SASICM 模型结构

4.2.2 Embedding 层

本节选取 GloVe[45] 和 BERT[9] 作为初始编码层。其中 GloVe 作为编码层在 CSD-数据集上进行训练得到，在网络中进行微调 (Fine-tuning)；BERT 作为编码层是采用了 Turc 等人[75] 在其文章中所使用预训练的模型，在网络中进行微调。除此之外，本文在编码侧，还增加了绝对位置编码，采用随机初始化的方式赋予初值。其中本文根据经验性，使用 GloVe 作为词向量训练方式的编码维度为 300，使用 BERT 作为词向量训练方式的编码维度为 768，位置编码的维度为 50。

4.2.3 强化注意力层

强化注意力层用于最大化重要词和不重要词之间的差异，并且让不同重要性之间的差异加大。强化注意力的实现是基于自注意力机制的：根据点积获取两个 Token t_i 和 t_j 之间的相似度，然后使用 Softmax 函数对 Token t_i 和其余任意 Token 之间的相似度做平滑，得到注意力分布 att_i。对 att_i 的某一项 k，若其低

于值 $b_{i,j}$，则将其置为 0，否则对其乘上一个缩放因子 $\theta_{i,j}$。将处理后的注意力分布通过 Softmax 函数，重新缩放回加和为 1，以保持 att 是一个分布的概念。公式4-1到4-6展示了如何计算强化注意力的过程。

$$s_{i,j} = \frac{\omega_q q_i \cdot (\omega_k k_j)^T}{\sqrt{d_h}} \tag{4-1}$$

$$s_i = [s_{i,0}, s_{i,1}, \cdots, s_{i,n-1}] \tag{4-2}$$

$$att_i = \text{Softmax}(s_i) \tag{4-3}$$

$$scaleds_i = \theta_i \cdot (\text{ReLU}(att_i - b_i)) \tag{4-4}$$

$$scaledatt_i = \text{Softmax}(scaleds_i) \tag{4-5}$$

$$r_i = scaledatt_i \cdot \omega_v v_j, \tag{4-6}$$

其中 $\omega_q, \omega_k, \omega_v \in \mathbb{R}^{d_e \times d_h}$ 是 q_i, k_j, v_j 的映射参数，d_e 是词向量的维度大小，d_h 是隐层维度，q_i, k_i, v_i 代表第 i 个位置的表征，三者于本文中相同。$r_i \in \mathbb{R}^{d_h}$。当 $x > 0$ 时，$\text{ReLU}(x) = x$，否则 $\text{ReLU}(x) = 0$。$\theta_i, b_i \in \mathbb{R}^l$ 都是一个实数值参数，l 是输入的序列长度。最后，带有注意力的句子序列表征 r_{fa} 为 r_i 之和：

$$r_{fa} = \sum_{i=1}^{l} r_i. \tag{4-7}$$

4.2.4 双向长短期记忆单元

长短期记忆网络 (Long Short-Term Memory, LSTM)[74] 是 Hochreiter 等人在 1997 年提出的一种网络结构，其适用于建模具有序列结构的数据类型。它通过设定神经元状态和输入门、输出门、遗忘门等门结构，使得每个前序时间点的信息都可以向后传递。与传统的 RNN[76] 相比，LSTM 可以缓解梯度消失和梯度爆炸问题，适用于建模较长的序列的结构。其基本结构如图4-2所示。其中 c_i 是神经元状态，它在当前神经元中更新信息，然后将其向后传递，σ 为 sigmoid 函数。其计算方式如公式4-8到4-12所示。公式4-9、4-8和4-10分别代表的是输入门、遗忘门和输出门，它们结构几乎完全一致。公式4-11和4-12展示的是神经元状态更新的方式。它融合了4-11所计算的候选状态、前序时刻的神经元状态和当前输入状态等信息，统一用于更新当前的

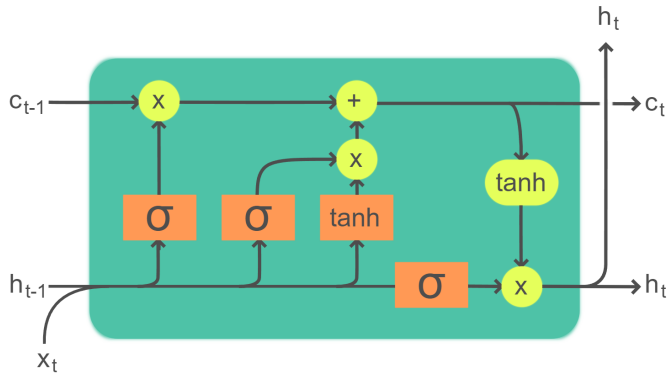


图 4-2: LSTM 的结构

神经元状态，并将之向后传递。

$$y_f(t) = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{4-8}$$

$$y_i(t) = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{4-9}$$

$$y_o(t) = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{4-10}$$

$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x] + b_c) \tag{4-11}$$

$$C_t = y_f(t) \times C_{t-1} + y_i(t) \times \tilde{C}_t, \tag{4-12}$$

其中，$\cdot$ 代表矩阵乘法，$\times$ 代表按位相乘，[] 代表拼接操作，$W_f, W_i, W_o, W_c \in \mathbb{R}^{(d_h+d_e)\times d_h}$ 是可学习的权重参数，$b_f, b_i, b_o, b_c \in \mathbb{R}^{d_h}$ 是可学习的偏置参数，$\sigma(x) = \frac{1}{1+\exp(-x)}$。将前向过程的输出记为 h_{ff}，将后向过程的输出记为 h_{fb}，则 Bi-LSTM 的输出为 $h_c = [h_{ff}, h_{fb}]$，且 $h_c \in \mathbb{R}^{2d_h}$。

4.2.5 特征融合层

当获取到带有注意力的句子内部特征 r_{fa} 和带有句子序列特征的 h_c 之后，将二者拼接。为了能够使得特征充分融合，使用一个线性变换，在原空间中进行特征变换。具体过程如公式4-13所示。

$$r_f = W_{fc}^T[h_c, r_{fa}], \tag{4-13}$$

其中 $W_{fc} \in \mathbb{R}^{3d_h \times 3d_h}$ 为可学习的参数矩阵。

4.2.6 反讽、比喻预测层

为了使得每个任务都能够学习到各自的特征，我们将 r_f 通过两个线性变换分别映射到反讽和比喻的特征子空间中，其中这部分特征仅来源于评论内容，而不包括上下文内容。如公式4-14和所示。

$$r_{sarc} = W_{sarc}^T r_f + b_{sarc} \tag{4-14}$$

$$r_{meta} = W_{meta}^T r_f + b_{meta}, \tag{4-15}$$

其中 $W_{sarc}, W_{meta} \in \mathbb{R}^{3d_h \times d_{subspace}}$，$b_{sarc}, b_{meta} \in \mathbb{R}^{d_{subspace}}$。得到两个空间上的特征向量之后，如公式4-16和4-17所示使用 Softmax 作为激活函数对其进行预测。

$$y_{sarc} = \text{Softmax}(W_{sarcp}^T r_{sarc}) \tag{4-16}$$

$$y_{meta} = \text{Softmax}(W_{metap}^T r_{meta}), \tag{4-17}$$

其中 $W_{sarcp}, W_{metap} \in \mathbb{R}^{d_{subspace} \times 3}$。

4.2.7 潜台词语义抽取模块

我们使用 MLP 作为语义抽取模块单元，首先将来自于评论的特征映射到潜台词的特征空间中得到 r_{subt}，然后使用 MLP 将 r_{subt} 映射到语义空间中，得到 $r_{mean-comm}$。同样的，我们将 S 分别经过 strengthen attention 和 Bi-GRU 得到的特征向量进行拼接，经过同一个语义抽取模块得到具有上下文的语义特征 $r_{mean-cont}$。完整过程如公式4-18到4-20所示。

$$r_{subt} = W_{subt}^T r_f + b_{subt} \tag{4-18}$$

$$r_{mean-comm} = \text{MLP}(r_{subt}) \tag{4-19}$$

$$r_{mean-cont} = \text{MLP}(W_{cont}^T [h_{cont,c}, r_{cont,fa}] + b_{cont}), \tag{4-20}$$

其中 $h_{cont,c} \in \mathbb{R}^{2d_h}$ 为完整的序列 S 经过 Bi-GRU 后得到的特征向量，$r_{cont,fa} \in \mathbb{R}^{d_h}$ 为 S 带有注意力的特征向量，$W_{subt}, W_{content} \in \mathbb{R}^{3d_h \times d_{subspace}}$，$b_{cont} \in \mathbb{R}^{d_{subspace}}$ 为可学习参数。

4.2.8 潜台词预测模块

将仅根据评论得到的语义向量 $r_{mean-comm}$ 和根据完整序列得到的语义向量 $r_{mean-cont}$ 拼接，使用 MLP 和 Softmax 函数进行预测。

$$y_{subt} = \text{Softmax}\left(\text{MLP}([r_{mean-comm}, r_{mean-cont}])\right). \tag{4-21}$$

4.2.9 损失函数

我们的损失函数由以下几部分构构成：

- 预测损失 L_{pred} 由三部分构成，分别为比喻预测的损失、反讽预测的损失和潜台词预测的损失。每部分损失都为交叉熵 (CrossEntropy)。

$$L_{pred} = w_{sarc}L_{sarc} + w_{meta}L_{meta} + w_{subt}L_{subt} \tag{4-22}$$

$$L_{sarc} = \sum_i y_{i,sarc} \log \hat{y}_{i,sarc} \tag{4-23}$$

$$L_{meta} = \sum_i y_{i,meta} \log \hat{y}_{i,meta} \tag{4-24}$$

$$L_{subt} = \sum_i y_{i,subt} \log \hat{y}_{i,subt}, \tag{4-25}$$

  其中 $w_{sarc}, w_{meta}, w_{subt} \in \mathbb{R}$ 为权重因子。

- Strengthen Attention 参数的约束，为了达到 Strengthen Attention 的目标，其参数不是无约束的，其不重要性的阈值 b_i 不能为负数，否则将认为全部都重要；其次不能太大，否则将认为全部都不重要，因此需要将其约束在一定范围内。Strengthen Attention 的缩放权重 θ_i 如果小于 1，则其起到的效果是减小差异，让注意力值更加平滑，而这有悖初衷，因此也需要对 θ_i 加以约束。综上，本文设定 $b_{i,j}$ 的范围需要在 $[\underline{b}, \bar{b}]$ 之间，否则会产生损失，同样的，设定 θ_i 不能低于阈值 $\underline{\theta}$。

$$L_{cons} = \sum_i \text{ReLU}(\underline{b} - b_i) + \text{ReLU}(b_i - \bar{b}) + \text{ReLU}(\underline{\theta} - \theta_i). \tag{4-26}$$

- 正则项，为了避免过拟合，我们对所有线性映射层以及 MLP 层中的所有参数都施加 L_1 正则化，记为 L_{l1-reg}。

  与此同时，考虑到三个任务中都有不同程度的数据不平衡现象，本文对

L_{pred} 应用代价敏感学习[77]。本文将数据集分成训练集、验证集以及测试集，代价敏感学习的权重根据验证集的结果随着训练过程进行调整。从验证集可以得到三个类别的 f_1 值 $f_{1,-1}$、$f_{1,0}$ 和 $f_{1,1}$，对它们进行归一化之后，乘上一个系数 g，作为不同类别的权重。

$$w_i = g\left(1 - \frac{\exp(f_{1,i})}{\sum_{i\in\{-1,0,1\}}\exp(f_{1,i})}\right). \tag{4-27}$$

4.3 实验与分析

本节对章节4.2中所述方法进行详细的实现介绍，进行相关的对比试验验证 SASICM 的有效性以及进行相关的消融实验以验证提出方法的有效性。本文在实验部分探讨了以下四个问题：

- 单任务的潜台词识别方法：本文在实验部分对比了单任务情况下 SASICM 简化模型，并且将对比的方法也进行了相关的操作。该设置的主要目的是对比多任务方法在我们的模型和问题上的效果与作用。
- 多任务的潜台词识别方法：本文对比衡量了 SASICM 和其余多任务的架构的表现。本文进行了双任务和三任务的模型在潜台词识别任务上的效果。该设置的主要目的是评估 SASICM 在潜台词识别任务上的优越性，同时确立本文所提的 SASICM 更适合作为一个新的基线模型 (Baseline)。
- 不同的词向量表征方法对于潜台词识别的影响：本文使用 GloVe[45] 和 BERT[75] 作为两种词向量表征的方式，分别在我们的任务上进行实现，对应的模型分别标记成 $SASICM_g$ 和 $SASICM_{BERT}$。该设置可以让读者直观的比较不同方式的词向量表征方法在潜台词任务上的差异。
- 人的表现：本文将同样的评估方法应用在人的标注结果上，记作 HP。该设置的目的是衡量我们的任务是否可做，并且给潜台词任务提供一个上限 (Upper Bound)。

4.3.1 基线模型的选取

本节主要介绍了在后文的实验中，为潜台词检测任务所选取的对比基线模型。

- **Bag-Of-Word + 经典的分类器**由于潜台词任务的隐晦性，其难以提取有效的特征，同时考虑人力成本，对于经典的分类器，本文采用词袋 (Bag-Of-

Word) 作为特征，使用 SVM、LR、NB、DT 和最大熵 (Maximum Entropy) 模型[78] 作为分类器对是否有潜台词进行判断。

- **BTM-Based Model** BTM 是一个二元任务的分类模型[55]。BTM 同时学习判断态度是否积极以及是否具有反讽，并且将模型对反讽的判断结果作为新特征用以加强观点态度的判断。在本文实验中，BTM 的两个任务设计成两组：1. 潜台词检测和反讽判断，对应的模型记成 $BTMSS$；2. 潜台词检测和比喻判断，对应的模型记成 $BTMSM$。同时，为了和 SASICM 的三任务模式对齐，本文还将 BTM 扩展成了三任务的模式，记成 $BTM3$；为了和单任务模式对齐，本文将 BTM 的模型降成了单任务的模型，其中潜台词检测对应的模型记成 $BTMSubt$，反讽判断对应的模型记成 $BTMSarc$，比喻判断对应的模型记成 $BTMMeta$。

- **MIARN-Based Model** MIARN[46] 是一个为了反讽识别任务设计的单任务模型架构。MIARN 设计了 Intra-Attention 模块用以更好的捕捉句子内部重要性特征，并且将序列特征和 Intra-Attention 特征同步抽取，而不采用堆叠的结构进行设计，避免了每个表征的相似度在经过序列模型处理后趋于相近的问题。类似于 BTM，潜台词检测对应的模型记成 $MIARNSubt$，反讽判断对应的模型记成 $MIARNSarc$，比喻判断对应的模型记成 $MIARNMeta$。本文也将 MIARN 采用 SASICM 一样的方式扩展成两组双任务模型：1. 潜台词检测和反讽判断，对应的模型记成 $MIARNSS$；2. 潜台词检测和比喻判断，对应的模型记成 $MIARNSM$；和采用 SASICM 一样的方式扩展成三任务模型，记作 $MIARN3$。

- **BERT+ 全连接神经网络** BERT[75] 在语言理解相关的任务上具有非常好的效果，其可以作为一个强对比模型，用以验证我们的方法是否有效。本文参照原论文所说的方式，将其输出直接使用全连接分类神经网络用以学习一个分类任务，进行整个模型的微调。类似于上文所述，将 BERT 相关的模型记成 $BERT3$、$BERTSS$、$BERTSM$、$BERTSubt$、$BERTMeta$ 和 $BERTSarc$。表 4–4的 $BERT + FF$ 即为 $BERTSubt$。

- **GBP** 为了证明我们的模型和对比模型的有效性，本文根据类别进行随机猜测结果 (Guess By Probability, GBP)，并将其用以作为一组对比，在对比实验中一同展示。

4.3.2 实验细节

本节旨在对本章涉及的模型细节进行描述。为了减缓模型过拟合的问题，SASICM 对需要分析的文本内容 W 和完整的输入序列 S 都采用了 dropout 技术，并且赋予了不同的 dropout 的值。除此之外，SASICM 还使用了交叉验证 (Cross Validation) 的方法进一步减缓这个问题，最终的呈现结果为交叉验证的平均值。衡量的方法，我们经验性的采用分类的衡量指标 F_1 Score 进行评估。在我们的实验中，我们使用单张 GPU(GTX 1080Ti) 进行训练，实验中涉及的超参数，如表4-1所示。

表 4-1: 超参数表

Embedding Size		GloVe	300	BERT	768	Positional	50
Meaning Subspace Size		100		$\underline{b}$	3e-3	$\bar{b}$	1
learning rate		1e-2		warm up step	50	θ	5
Random seed	42	batch size	12	patience for early stop		10	
dropout rate for W	0.2	dropout rate for S	0.4	cross validation fold		5	

4.3.3 对比实验及其分析

本节将经典的机器学习方法和三种基于神经网络的方法与 SASICM 进行对比，考虑到对比实验的公平性，将对比实验分为三组，分别为在单任务上的对比实验结果、双任务模型上的对比实验结果和三任务模型的对比实验结果。其中三任务为 SASICM 完整的模型，双任务的 SASICM 基于三任务削减了其中一个任务的分支得到，单任务仅留下潜台词任务这一个分支。类似的操作，也同样的应用于对比的三种神经网络方法。经典的机器学习任务由于其只有一个参数，其本身就限定了只做一个任务的分类，因而无法扩展，仅在单任务对比试验上进行呈现。

4.3.3.1 与基线模型的指标对比

表4-2展示的是三任务上的对比实验结果，表4-3展示的是双任务上的对比结果，表4-4展示的是单任务上的对比结果。

在潜台词任务上，$SASICM$ 在三任务上的正确率略高于其它模型，与 MIARN3 的表现相近，在双任务的表现低于其他神经网络模型，在单任务设定

表 4-2: 三任务模型的实验结果。其中 “p” 为准确率，“r” 召回率。下划线所表示的为我们的模型 (SASICM)，$*_m$ and $*_s$ 分别代表反讽 (sarcasm) 和比喻 (metaphor) 的结果。

Model	Subtext Task				Metaphor Task				Sarcasm Task			
	p(%)	r(%)	F_1(%)	acc(%)	p_m(%)	r_m(%)	F_{1_m}(%)	acc_m(%)	p_s(%)	r_s(%)	F_{1_s}(%)	acc_s(%)
SASICM$_g$	67.57	68.52	67.04	68.52	83.72	91.50	87.43	91.50	84.38	91.86	87.96	91.86
SASICM$_{BERT}$	66.35	66.38	65.89	66.38	84.73	88.02	86.28	88.02	85.57	90.02	87.70	90.02
MIARN3	62.89	68.04	63.80	68.04	83.34	91.12	86.73	91.12	82.51	90.84	86.48	90.84
BTM3	63.24	67.33	63.08	67.33	84.25	90.26	86.89	90.26	82.62	91.08	86.43	91.08
BERT3	63.98	64.59	64.22	64.59	84.9	86.72	85.76	86.72	85.39	88.64	86.97	88.64
GBP	57.39	57.38	57.38	57.38	84.83	85.22	85.02	85.22	85.44	85.76	85.60	85.76
HP	81.05	76.82	78.20	76.82	92.20	89.54	88.65	89.54	93.01	92.90	92.89	92.89

表 4-3: 双任务框架的实验结果

Model	Subtext Task				Metaphor Task				Sarcasm Task			
	p(%)	r(%)	F_1(%)	acc(%)	p_m(%)	r_m(%)	F_{1_m}(%)	acc_m(%)	p_s(%)	r_s(%)	F_{1_s}(%)	acc_s(%)
SASICMSS	65.17	69.02	66.07	69.02	-	-	-	-	84.38	91.86	87.96	91.86
MIARNSS	62.15	71.61	63.94	71.61	-	-	-	-	85.33	91.14	87.87	91.14
BTMSS	61.66	71.31	62.52	71.31	-	-	-	-	84.98	**92.19**	87.82	**92.19**
BERTSS	61.97	72.09	64.41	72.09	-	-	-	-	84.98	**92.19**	87.82	**92.19**
SASICMSM	63.23	66.39	62.80	66.39	83.78	90.97	87.21	90.97	-	-	-	-
MIARNSM	61.75	70.65	63.68	70.65	84.62	91.27	87.62	91.27	-	-	-	-
BTMSM	61.57	71.71	62.27	71.71	84.35	91.72	87.85	91.72	-	-	-	-
BERTSM	61.97	72.02	64.40	72.02	84.32	91.82	87.91	91.82	-	-	-	-

下正确率略低于其余模型，但是比朴素贝叶斯和决策树高。但由于数据存在不平衡性，而正确率存在偏向多数类的缺点。F_1 值同时兼顾了准确率和召回率，即考虑到了数据不平衡问题带来的衡量偏差，因此，本文的衡量指标中，F_1 值比正确率更加重要。

表 4-2中，SASICM$_g$ 代表使用 GloVe[45] 作为词向量方式的模型，SASICM$_{BERT}$ 代表使用 BERT[9] 作为词向量方式的模型，SASICM$_{BERT}$ 在潜台词检测上的 F_1 值比 SASICM$_g$ 低将近 1%，但是在精确率上比 SASICM$_g$ 高 1.23%，SASICM$_{BERT}$ 在召回率上比 SASICM$_g$ 低 3%。并且，在训练速度和推理速度上，SASICM$_{BERT}$ 比 SASICM$_g$ 要慢将近 1 倍，并且耗费的内存要大将近两倍。考虑到实验效果、推理时间、训练的收敛速度，本文将 SASICM$_g$ 作为任务的新 baseline。SASICM$_g$ 关于潜台词检测的 F_1 值比 BTM3 高 3.15%，正确率比之高 1.6%；在反讽和比喻两个子任务上的 F_1 也要比 BTM3 表现更优异。SASICM$_g$ 和 BTMSS 相比，关于潜台词检测的 F_1 值比 BTMSS 高 3%，但是正确率略有不足；SASICM$_g$ 对于反讽判断的 F_1 值与 BTMSS 相当，但是在正确率上略低于其。SASICM$_g$ 和 BTMSM 相比在潜台词上的差异与 BTMSS 相当，但是在隐

喻上的效果要略低于 BTMSM，但是差异不大。SASICM$_g$ 和 BTMSubt 相比，在 F_1 上提升将近 4%，但正确率比之低 3%。对比可知，SASICM$_g$ 比 BTM 更适合处理偏类问题。类似的，SASICM$_g$ 与 MIARN3、MIARNSS、MIARNSM 和 MIARNSubt 相比，也更适合偏类问题。在三任务上，SASICM$_g$ 在两个指标上都比 MIARN3 更好，但是对比单任务和双任务，SASICM$_g$ 都在 F_1 上更有优势，但是正确率上略低。同时，对比 BTM 和 MIARN 系列的模型，MIARN 可以取得比 BTM 更好的 F_1 值，同时正确率略低于 BTM。基于 BERT+ 全连接层 (FF) 和 SASICM$_g$ 的方法相比，可以取得除了 SASICM 之外最好的结果，但是其训练速度和推理速度要比慢很多，且无论多任务单任务，BERT+ 全连接层 (FF) 的结果几乎无变化。可见在偏类问题上，对 SASICM 和其余三个基于深度学习的模型进行排序，可以简单得到：SASICM$_g$ > BERT + FF > MIARN > BTM。

由于经典方法的特殊性，无法直接使用其做多任务模型，因而，对于这类方法，仅做了表4-4中所示实验。从结果中可得，绝大部分经典方法的结果与基于神经网络的方法的相比，相去甚远。尤其是基于朴素贝叶斯的方法，F_1 值差距很大。但是基于朴素贝叶斯的方法可以取得比其余传统方法更高的准确度，这说明朴素贝叶斯学习到了部分潜台词的特征，但是存在过拟合的现象，只要有这部分特征就为潜台词，反之则无。且这类特征比较简单，没有考虑到各个特征之间的组合和交互。决策树的方法在经典的方法中表现最好，可以和 BTM 的表现相当，但是其在准确度上表现较差。

表 4-4: 单任务框架的实验结果

Model	Subtext Task				Metaphor Task				Sarcasm Task			
	p(%)	r(%)	F_1(%)	acc(%)	p_m(%)	r_m(%)	F_{1_m}(%)	acc_m(%)	p_s(%)	r_s(%)	F_{1_s}(%)	acc_s(%)
SASICMSubt	63.80	69.99	66.56	69.99	-	-	-	-	-	-	-	-
MIARNSubt	61.70	70.56	63.64	70.56	-	-	-	-	-	-	-	-
BTMSubt	62.04	71.96	62.59	71.96	-	-	-	-	-	-	-	-
BERT+FF	61.98	72.10	64.41	72.10	-	-	-	-	-	-	-	-
SVM	60.67	72.00	60.60	72.00	-	-	-	-	-	-	-	-
LR	55.93	72.05	60.44	72.05	-	-	-	-	-	-	-	-
MEC	51.98	72.10	60.40	72.10	-	-	-	-	-	-	-	-
NB	61.14	11.50	13.35	11.50	-	-	-	-	-	-	-	-
DT	62.19	66.62	63.09	66.62	-	-	-	-	-	-	-	-

对比每个模型从单任务到双任务，再到三任务的衡量指标变化情况，发现基于 BERT+FF 的模型在响应任务上基本没什么变化，增加一个任务，就只学习了相关任务的特征，原有任务并不会受到影响。产生这种现象的原因，从模

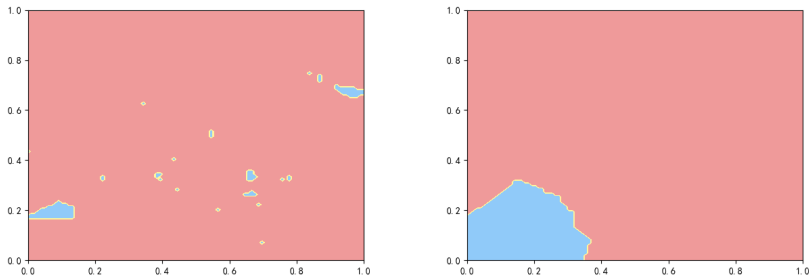
型和数据的角度，本文认为，大模型可以比较容易的学到较少数据上的所有特征，并且产生一定的过拟合，通过 BERT 出来的特征是非常多的，每个下游任务都只是从特征中抽取相关的部分进行分类。基于 BTM 的模型，在双任务上的效果都没什么提升，甚至有轻微的下降，但是在三任务上却可以取得更好一些的效果。这个说明，只是单纯的增加一个任务，影响比较小，会让模型从原有的任务学习到稍微偏向新任务的特征。而多个新任务使得潜台词特征的学习更具有指导性，而不仅仅是提供一个特征。基于 MIARN 的系列模型中，MIARNSS 在潜台词上取得了最好的效果，MIARNSubt 效果最差，MIARN3 的效果居于 MIARNSM 和 MIARNSS 之间，略高于单任务的模型。不同子任务的特征对于潜台词识别的增幅是不同的。从准确率和召回率上分析，不同的子任务对于 MIARN 而言，可以提升潜台词的识别准确率和召回率。但当子任务同时存在时，却会对召回率产生影响，即存在多个子任务时，MIARN 倾向于对不同的子任务产生的特征取交集用来判断潜台词，使得判断标准变得更加严格，却有更大的置信度让识别出来的潜台词是正确的。

4.3.3.2 表征分析

本节对每个模型的倒数第二层表征进行分析。本文通过 t-SNE[6] 方法将表征从高维特征特征空间降维到二维特征空间上，呈现出一个一个的数据点。对数据点使用 K-nearest neighbors (KNN)[79] 算法进行分类。KNN 中的 K 本文指定为 20。由于 0 代表不确定，这在实际用途上的使用具有不确定性，因此我们将这部分类别去除，在图 4-3中，仅呈现出 1（具有潜台词）和-1（不具有潜台词）所处类别的分类边界，其中粉红色区域代表-1 类别，而蓝色区域代表 1 类别。直观上，如果数据表征可以使得同一类别的数据点相距较近，不同类别的数据点相聚较远，则可称之为好的数据表征。同时，如果该类别具有多个不同的模式，则该类在数据点的聚类结果中，应该呈现多点分布，或者同一分布区域应该呈现多角的不规则形态。由于不同的模型可能学习到潜台词的不同的特征，这会导致 t-SNE 降维后的位置完全不一样。因此，在对表征进行分析时，我们更加关注蓝色区域分布的大小和数量，以及它所呈现的形态，而不关注它所出现的位置。蓝色区域越大，说明它学习到了更加泛化的特征，可以识别出更多的数据点；蓝色区域分布越多，说明它学习到了更多的模式，可以对不同的模式都有效果。由此可得，蓝色区域越大分布越多都说明表征效果越好。

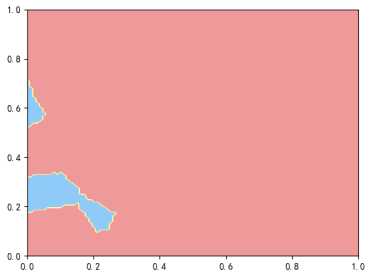
图 4-3(f)和 4-3(e)展现的是 SASICM 在不同词嵌入方式下学习到的倒数第

二层表征的聚类效果，观察可知，SASICM 的方法学习的表征聚类后的面积要比其余方法的大很多。且 SASICM$_g$ 学习到的表征聚类后的区域更多，而 SASICM$_{BERT}$ 学习到的特征则更为集中，却没那么发散。而基于 BERT+FF 的方法学习到的表征聚类面积，如图 4-3(a)所示，呈现出类很多的区域，但是每一块区域的面积都非常小，即基于 BERT+FF 的方法可以学习到很多的模式，但是几乎在每个模式上都产生了过拟合现象，而这也是 BERT+FF 在潜台词任务上效果不好的原因。相似的情况出现在了基于 BOW+ 经典机器学习方法的上，图 4-3(d)展示的是决策树学习的输入特征，可以看到，如果直接使用 BOW 作为词的表征，也会如 BERT 一般，可以涵盖大多的模式，但是每个模式上都几

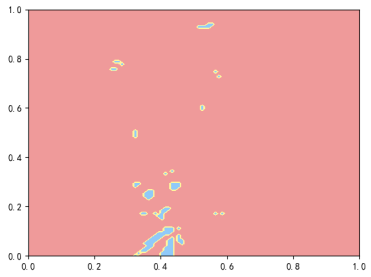


(a) BERT+FF 的表征

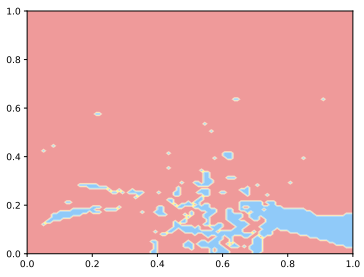
(b) BTM 的表征



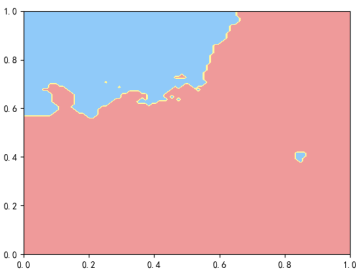
(c) MIARN 的表征



(d) BOW 的表征



(e) SASICM$_g$ 的表征



(f) SASICM$_{BERT}$ 的表征

图 4-3: 不同模型的表征示例。表征都来源于不同模型的倒数第二层，使用 t-SNE[6] 对倒数第二层的隐向量做降维，并且使用 KNN 做聚类所得。

乎产生过拟合的现象。而基于 BTM 的方法，如图4-3(b)所示，表征聚类后的面积仅次于 SASICM，但是仅有一块蓝色区域，这说明基于 BTM 的方法仅学习到了一类相似特征，对其余模式的识别效果几乎没有，这是 BTM 效果不好的一个原因。基于 MIARN 的方法，如图4-3(c)所示，它的蓝色区域面积虽然没有 BTM 大，但是它学习到的模式可以可以认为有三个，比 BTM 要更多。但是和 SASICM 相比，蓝色区域面积不够大，且模式没有 SASICM 多。

4.3.4 消融实验及其分析

从表4-4、4-3和4-2可以看到 SASICM$_g$ 从单任务、双任务到三任务之间的指标变化。从评估指标的变化情况可以得到，并不是任意多任务都会带来指标的提升，单任务上的 SASICM 效果比双任务好。但是双任务也并非一无是处，其中反讽任务加入可以提高对潜台词判别的准确率。最好的结果是三任务上的效果，三任务上，以稍微降低一些召回率的代价，可以大幅度的提高准确率。由此，我们可以判断，当只有两个任务的时候，模型同时学习了潜台词和辅助任务的特征，由于辅助任务比较简单，此时单一辅助任务对模型的影响较大；当有三个任务的时候，模型在三个人任务之中进行权衡，此时辅助任务对模型的影响将不再是单一方向的引导，而是多个方向的，此时不同方向的特征信息可以对潜台词的判断起正面作用。

除了表4-4到表4-2所示的实验之外，本文还增加了其余几组消融实验以验证模型设计的合理性。

为什么不是用 Self-Attention：自注意力机制 (Self-Attention) 在自然语言处理应用中使用非常广泛，比如 Transformer[59]。在大多数情况下，自注意力机制在语言理解方面的任务中，确实可以取得比较好的效果，因此本节也对其进行分析，将本文所提的 Strengthen Attention 替换成 Self-Attention，进行对比。其实验结果如表4-5 SASICMSA 所示。

完全数据驱动的 Strengthen Attention 会不会更好：本文对提出的 Strengthen Attention 的参数添加了一些限制，这在一定程度上违背了数据驱动的原则，具有比较大的主观性和经验性。那么是否让 Strengthen Attention 完全数据驱动会有更好的效果？为此，本节通过实验来验证所添加的限制是合理的。其实验结果如表4-5 SASICMWC 所示。

为什么不对 Strengthen Attention 和 LSTM 结构采用堆叠的方式进行组合：Tay 等人[46] 提出：在序列结构之后，再使用 Attention 方式进行特征建模，其每

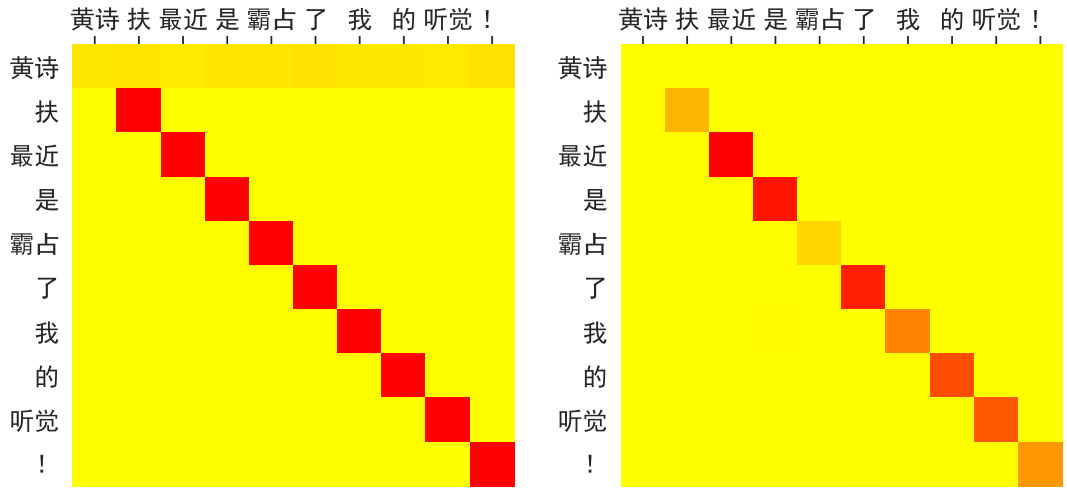
个 Token 的表征会变得非常相似。本节对此，在 SASICM 的基础之上进行验证。其实验结果，如表 4-5 SASICMSt 所示。除此之外，本节还探究了 GRU 和 LSTM 之间的区别以及在建模中使用单向的 LSTM 等问题。

Strengthen Attention 比 Self-Attention 好的原因：表 4-5中可以看到将 Strengthen

表 4-5: 消融实验结果

Model	Subtext Task				Metaphor Task				Sarcasm Task			
	p(%)	r(%)	F_1(%)	acc(%)	p_m(%)	r_m(%)	F_{1_m}(%)	acc_m(%)	p_s(%)	r_s(%)	F_{1_s}(%)	acc_s(%)
SASICMSt	67.68	69.13	66.61	69.13	83.72	91.50	87.43	91.50	84.38	91.85	87.96	91.85
SASICMG	65.12	68.93	66.23	68.93	83.72	91.50	87.43	91.50	84.38	91.86	87.96	91.86
SASICMWC	64.46	70.07	66.25	70.07	83.32	91.69	87.30	91.69	84.38	91.85	87.95	91.85
SASICMSL	66.07	68.53	66.28	68.53	83.72	91.50	87.43	91.50	84.38	91.85	87.96	91.85
SASICMSA	65.47	67.94	65.16	67.94	84.58	91.00	87.40	91.00	84.68	89.35	86.82	89.35

Attention 替换成 Self-Attention 之后，效果降低了将近 2%。可见 Strengthen Attention 比 Self-Attention 在潜台词任务上效果更好。以“黄诗扶最近是霸占了我的听觉！”为例，将每个字的 Self-Attention 注意力值和 Strengthen Attention 注意力值进行比较。两个不同注意力值可视化结果如图 4-4所示。从两张图中可以直观的看出，Self-Attention 主要将注意力都放在了自己的身上，而且没有什么区分度；而 Strengthen Attention 则不同，区分度较为明显。



(a) Self-Attention 的注意力值　　(b) Strengthen Attention 的注意力值

图 4-4: 自注意力机制 (Self-Attention) 和强化注意力机制 (Strengthen Attention) 的对比

完全数据驱动的 Strengthen Attention 的问题：完全数据驱动的 Strengthen Attention 结果如表 4-5所示，可见如果没有限制的话，效果要比加以限制的低 0.8%。对没有限制的 Strengthen Attention 参数进行分析，发现自动化学得的缩

放权重 $\theta \in (-1, 1)$，阈值 b 接近于 0；而加了限制的 Strengthen Attention 的缩放权重 $\theta \in (0, 10)$ 之间，其中绝大部分数值分布在 $(1, 5)$ 之间，阈值 $b > 3e - 3$。

堆叠结构的问题：堆叠结构的结果如表4-5所示，效果低了将近 0.5%，取 Strengthen Attention 的结果进行分析，发现堆叠结构的 Strengthen Attention 之后的词表征相似度平均约为 0.96 左右，而非堆叠结构的词表征相似度平均约为 0.87。Tay 等人[46] 所提的问题确实存在，且相似度表征确实会影响最终的实验结果。

其它：本节还探究了使用 GRU 代替 LSTM 的实验效果，如表4-5 SASICMG 所示，发现使用 GRU 实验效果要比 LSTM 低将近 0.8%，时间要比 LSTM 快 12.5%（GRU 使用 3.5 小时完成整个实验，而 LSTM 使用 4 小时）。虽然 GRU 提高了运算的速度，但是效果却有比较明显下降。同时本节还探究了使用单向结构的实验效果，如表4-5 SASICMSL 所示，发现效果下降也将近 0.8%，可见对于语言理解类的任务，双向建模比单向建模更好。

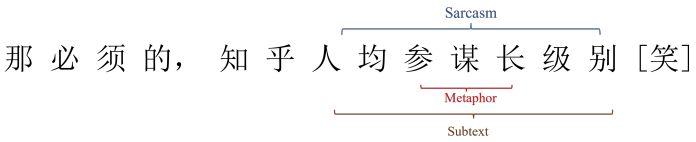
4.4 本章小结

在本章中，我们对潜台词检测这个任务做了更加明确地辨析，同时提出了一种双塔结构的多任务模型——SASICM。SASICM 通过对序列特征和序列内部特征分别建模得到了不同类型的特征。SASICM 同时也对目标序列和完整的上下文序列采用相同的模型分别建模，得到了不同信息下的特征。简单的通过不同的分支，实现了对反讽和比喻的识别，通过语义抽取模块对是否具有潜台词进行判断。且 SASICM 的并行结构，利于 GPU 对其进行加速，计算速度快。最后，本章还对 SASICM 进行了一系列对比试验和消融实验，并对其进行了详细的分析。

第五章　片段级分类任务：潜台词与实体片段识别

上一章中我们提出了句子级别的分类任务——潜台词检测。在实际的生产实践以及分析中，仅知道该句具有潜台词，是不够的，更需要知道它哪个片段具有潜台词以及是什么样的潜台词。与该任务相类似的任务如相关工作中所述有平展类型的序列标注任务和嵌套类型的序列标注任务（嵌套命名实体识别）。从本质上，三者的目标都是识别出完整文本中的某个片段，并且判断其是否属于特定类型，同时都可以建模成对每个词或者 Token 进行的分类任务。本章内容旨在对该类任务既有算法的一些不足，进行改进，并且将本文所提算法应用于潜台词的内容片段识别和嵌套的命名实体两个任务上。

5.1　内容片段识别任务的特点



(a) 样例 1

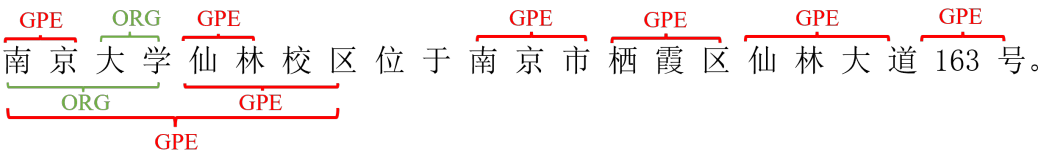
(b) 样例 2

(c) 样例 3

图 5-1: 潜台词片段识别的三个样例

图 5-1所示为潜台词片段识别的三个例子，分别涵盖了图 5-1(a)所示，潜台词内容中包含反讽内容片段和比喻内容片段；如图 5-1(b)所示，反讽内容中包含潜台词内容；和如图 5-1(c)所示，三个内容互不包含的情况。我们将同时具有反讽或者比喻内容的潜台词称为反讽类型或者比喻类型的潜台词，其余称为普通类型的潜台词。

(a) 样例 1



(b) 样例 2

图 5-2: 嵌套命名实体识别的两个样例

嵌套的命名实体识别任务，如图 5-2所示，类似于潜台词片段识别任务，也存在多种类型的相互嵌套或者不嵌套的情况。对于不嵌套的实体识别任务即退化为平展类型的序列标注任务，显然，嵌套的命名实体识别任务更具泛化性和通用性。对于图 5-1和图 5-2所示任务，可对其进行抽象化为：给定一个序列 $S = \{w_1, w_2, \cdots, w_n\}$，识别 $S_{sub} = \{w_i, \cdots, w_j\}$ 的所属类型。

5.1.1 现有工作的不足

如章节 2.3.1所述，平展类型的序列标注算法可将每个 Token 分类到一种类别中，但是其缺点是无法处理嵌套的情况，需要对其扩展才能得以适用于嵌套类型。其中一种扩展方法为设定阈值，当 Token 的某一个分类结果高于阈值时，则保留其标签类型。这种扩展方法有其不可避免的两个缺陷：1. 对于保留的标签类型，其类型组合产生的多余负样本，在推理阶段难以去除。以“ABCDEFG” 为例，B 保留了 “B-1, B-2” 类型，C 保留了 “E-2, I-2, I-1” 类型，D 保留了 “E-1, E-2” 类型，则可能产生的子序列为 “BC” 为类型 2，“BCD” 为类型 2，“BCD” 为类型 1。而正样本只有类型 2“BC” 和类型 1“BCD” 时，属于类型 1 的 “BCD” 序列无法被有效去除。2. 阈值难以选取。由于没有一个可信的方法用于选取阈值，因此阈值的选取通常需要通过搜索进行实现，而这非常耗时耗

力。另一种扩展方法为基于多任务框架解决嵌套问题，每种类型使用一个子任务进行预测，该种扩展方法可以解决大部分的嵌套情况，但当有同种类型嵌套时依旧无法避免只能识别出其中一种的情况。

专门针对于嵌套类型的序列标注识别算法比平展类型的序列标注算法进行扩展，更具有有效性和通用性。然而，这类算法也有其局限性。基于边界预测的算法，存在错误传播的问题，即当边界识别错误时其会影响到下游的分类结果。基于阅读理解的算法，其查询语句的构造是复杂且难以泛化的。基于两段式的方法，其效果最好，但是其预处理较为复杂，同时在第一阶段的时间和空间复杂度高。对于长度为 L 的序列，设其枚举的最大窗口为 M，则基于两段式的方法在第一阶段需要枚举的次数为 $(2L-M)*M/2$，为 $O(ML)$ 的时间和空间复杂度。

本章所提基于“中心度”的方法主要目的是为了解决基于两段式方法预处理复杂且在应用时时间和空间复杂度较高的问题。

5.1.2 中心度及其优势

本文定义一个特殊类型的文本片段中心为它所处位置的序列中心，即假设其位置为 $p=\{p_i,p_{i+1},\cdots,p_j\}$，则其中心为 $center=\frac{\sum_i^j p_i}{j-i+1}$。本文定义片段中某个词的中心度为偏离片段中心的程度，即当序列上的点偏离其中心越远时，其在所属片段中的重要性下降程度越多。其计算方法如下：

$$c_i = 1 - \frac{|p_i - center|}{j-i+1}. \tag{5-1}$$

中心度如何解决两段式方法时间和空间复杂度较高的问题？本文以图5-3为例进行说明，该例子来源于开源的生物数据集 GENIA。对于两段式的方法而言，需要枚举所有可能的 Span，在图示例子中，Span 长度的最小范围为 1$\tilde{4}$，如此才能覆盖到图示例子中所有可能的实体。在最小范围的条件下，两段式结构需要枚举的 Span 数目为：（24 + 23 + 22 + 21）共 90 个。然而，若是借用中心度，则可以将该数目大为减少。中心度首先可以预测出每个词所对应的重要性程度。根据中心度的定义，中心度的理论最小值为 0.5，出于模型的误差考虑，可以适当放宽这个限制。根据中心度可以最大程度的去除对于无关位置区域的枚举。在图示例子中，理想情况下，仅需要对红色字体所在区域进行枚举即可，则其基于中心度所需要枚举的 Span 数目为：（4+3+2+1）+（2+1）共

- Induction of Jurkat leukemic T cells with phorbol 12-myristate 13-acetate and ionomycin did not affect the level of FKBP mRNA .

图 5-3: 为什么需要中心度的样例

13 个。可见，基于中心度的枚举可以极大的提高枚举的效率和减小时间和空间损耗。

中心度为何比 BIEO 的边界预测更好？中心度相比于 BIEO 的边界预测，其通过数值，软性引入了边界信息和偏离特殊序列中心的程度。数值越接近 0.5，则该 Token 所处位置越接近边界，且越原理特殊序列的中心。而 BIEO 在每个位置上仅能看到自己的信息，除了 B 与 E 以外，并没有任何边界信息包含其中。

为了验证中心度的有效性，本文将中心度的思想应用在两个任务上，分别为本文所提出的潜台词片段识别和嵌套命名实体识别，并且本文也根据两个任务的不同特点，设计了相应的解决方案。

5.2 基于中心度的潜台词片段识别

潜台词片段识别需要对同一个句子识别其比喻片段、反讽片段和普通的潜台词片段。三种不同的语义片段，通过同一个模型进行抽取。以“那必须的，知乎人均参谋长级别 [笑]”为例，如图 5-1(a)所示，它所处上下文和背景为“之前在知乎上刷关于卡大佐的问题，记得有条评论是:“哪怕卡大佐在知乎上随便拉个人去当参谋，也不会死得这么难看”[笑哭][笑哭]”。该句表现出来隐藏意思为“讽刺知乎上面的人都自认为自己很厉害，而评论者却不这么看”，其中“人均参谋长级别”既表达了“评论者不这么看知乎人均参谋长”这一讽刺的含义，同时“参谋长”又用来比喻“聪明的人”。类似的，如图 5-1(b)，在例子“好家伙，起码九年”中，其所处的上下文为“一个月一期，100 期可以追几年了”，背景知识为“BiliBili 的视频”，该句整体表现出反讽的意味，“起码九年”具有潜台词，明面上表示“时间久”，实则说明“更新拖沓”。本节的主要内容为设计一个基于中心度的解决方案，使之可以抽取出比喻、反讽、潜台词三者在原文字中对应的序列，并且可以适应嵌

套、相交和不相交的情况。其严格数学表达为：设 $type_i \in \{比喻, 反讽, 潜台词\}$ 为分类目标，$S = \{< cls >, w_1, w_2, \cdots, w_m, < sep >, c_1, \cdots, c_n\}$ 为输入序列，$h_i \in \mathcal{R}$ 为分类目标子序列起始位置，t_i 为分类目标子序列的终止位置，则潜台词片段识别的目标是学习一个函数 $F(x)$，使得 F 可以产生下述三元组集合 $T = \{(h_1, t_1, type_1), (h_2, t_2, type_2), \cdots, (h_m, t_m, type_m)\}$，其中 $t_i, h_j \in (1, m)$，$S_W = \{w_1, w_2, \cdots, w_m\}$ 为目标分析序列，$S_C = \{c_1, \cdots, c_n\}$ 为目标分析序列所处的上下文序列。

潜台词片段识别具有一定的特殊性：1. 同类型的潜台词相互不嵌套；2. 每个句子所需要识别出来的内容片段少。基于此，本文提出多任务的模型框架，分别对比喻、反讽和普通的潜台词内容片段进行识别。如果对每个 Token 进行二分类，即识别其是否是起点或者终点，该方法训练与测试的正负样本比约为 1.9 : 100，可见其数据非常不平衡。如果将每个 Token 分类成 BIEO 中的一个标签，其正负样本比约为 9.4 : 100，可以减缓正负样本比不平衡的问题。使用中心度进行预测，除了和 BIEO 分类方法一样可以减缓正负样本比不平衡的问题之外，还可以在每个位置上软性地引入边界信息，因为从中心度的数值可以预

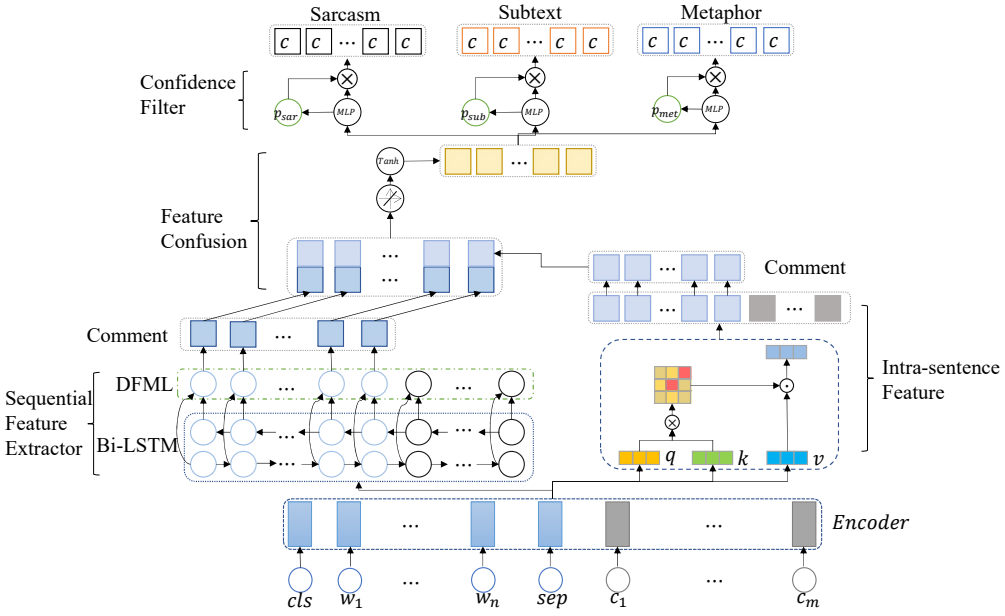


图 5-4: CBM 模型结构图

估其距离边界的程度。因此本文提出使用 “中心度” 作为预测目标，并基于此提出 “基于中心度的多任务内容片段识别模型 (**C**enterness **B**ased **M**ulti-task Content Segment Recognition Model, CBM)”。

5.2.1 模型结构

本节旨在对 CBM 模型做一个详细的介绍，其模型结构如图5-4所示。包括编码层、序列特征抽取模块、句内特征抽取模块、特征融合层、三个置信度预测层和三个内容片段识别层。

5.2.1.1 编码层

我们任务的输入如前文所述，为完整包含上下文的序列 $S = \{< cls >, w_1, w_2, \cdots, w_m, < sep >, c_1, c_2, \cdots, c_n\}$。在编码层本文使用 GloVe[45] 在我们的语料上进行训练，得到相关的词向量。使用 BERT[9] 直接在我们的任务上进行微调。同时，我们还增加随机初始化的绝对位置编码用以标记每个 Token 的位置。

5.2.1.2 序列特征抽取模块

序列特征抽取模块包含了两个部分，其一为上下文序列特征抽取，其二为深度上下文特征融合层。其中，序列特征抽取采用 LSTM 作为基本单元，对上文和下文的特征分别进行建模。通常而言，完整的上下文特征可以采用双向的 LSTM 作为基本单元，然后将上文和下文得到的隐层特征进行拼接得到。然而，这种方法对于上下文的特征融合而言，是比较粗糙的。为此，本文提出一种基于 LSTM 的深度上下文特征融合方法，用于对序列的特征进行建模，称之为深度特征融合层 (**D**eep **F**eature **M**ixture **L**ayer, DFML)。DFML 与通常做法不同之处在于，获取前后向的隐层向量后，不采用拼接的方式对上下文信息进行建模。针对位置 i，其计算了上文表征和下文表征每个维度的最大值与平均值，然后对最大值与平均值进行降维，达到深度特征融合的目的。其中最大值期望保持上下文特征的特性，而平均值期望保持上下文特征的共性。其中，第 i 个位置的上文建模的隐层特征记为 $\vec{h_i}$，下文建模的隐层特征记为 $\overleftarrow{h_i}$，则对于第

i 个位置而言，其特征融合过程如公式 5-2到 5-6所示。

$$\vec{h_i} = LSTM([e_0; \cdots ; e_i]) \tag{5-2}$$

$$\overleftarrow{h_i} = LSTM([e_{n+m+2}; \cdots ; e_i]). \tag{5-3}$$

$$h_{max,i} = MaxPooling([\vec{h_i}; \overleftarrow{h_i}]) \tag{5-4}$$

$$h_{mean,i} = MeanPooling([\vec{h_i}; \overleftarrow{h_i}]) \tag{5-5}$$

$$h_{full,i} = [h_{max,i}; h_{mean,i}] \tag{5-6}$$

$$h_{full} = [h_{full,0}; h_{full,1}; \cdots ; f_{full,n+m+2}], \tag{5-7}$$

其中，$e_i \in \mathbb{R}^{d_e}$ 为第 i 个位置的嵌入向量，d_e 为嵌入向量的特征维度，$MaxPooling$ 为按特征维度进行的最大池化，$MeanPooling$ 为按特征维度进行的平均池化，$\vec{h_i}, \overleftarrow{h_i}, h_{max,i}, h_{mean,i}, h_{full,i} \in \mathbb{R}^{d_e}$，$h_{full} \in \mathbb{R}^{(n+m+2)\times 2d_e}$。同时，为了让上下文特征可以更充分融合，将 h_{full} 通过非线性变换，映射到低维空间之中。

$$h^{'} = \sigma\left(MLP(h_{full})\right), \tag{5-8}$$

其中，$h^{'} \in \mathcal{R}^{(n+m+2)\times(d_e)}$，$d_e$ 为隐层特征的维度。MLP[80] 为多层感知机，本文通过多层感知机实现隐层特征从高维空间向低维空间的转换。$\sigma(x) = \frac{1}{1+\exp(-x)}$ 为 sigmoid 激活函数。

5.2.1.3 句内特征抽取层

本文使用原始的 Self-Attention 方法构建句内特征。计算其余所有位置关于位置 i 上的关注度，并以之为权重对所有位置的表征进行加权求和，得到位置 i 上的句内特征 $h_{I,i}$，从而得到整个序列的句内特征表示为 $h_I = [h_{I,1}; h_{I,2}; h_{I,n}]$。

5.2.1.4 特征融合层

得到了序列特征 $h^{'}$ 和句内特征 h_I 后，仅保留其中和目标语句相关的部分，记为 $h_w^{'}$ 和 $h_{w,I}$，且 $h_w^{'} \in \mathcal{R}^{m\times d_e}$，$h_{w,I} \in \mathcal{R}^{m\times d_I}$。我们将二者拼接后，通过非线性变换对两种不同的特征进行融合。

$$h_m = [h_w^{'}; h_{w,I}] \tag{5-9}$$

$$h_{mt} = W_m^T Tanh(h_m) + b_m, \tag{5-10}$$

其中，$h_m \in \mathcal{R}^{m\times(d_l+d_e)}$，$W_m \in \mathcal{R}^{(d_l+d_e)\times d_m}$ 为可学习参数，d_m 为变换后的特征空间维度，$b_m \in \mathcal{R}^{m\times d_m}$ 为可学习的偏置参数。$Tanh(x) = \frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}$ 为非线性变换的函数。

5.2.1.5 中心度预测层

得到融合后的特征 h_{mt}，将之分别映射到反讽、比喻和普通潜台词的特征空间中，得到其各自的特征。在每个特征空间中通过简单的线性回归，预测每个 Token 所对应的中心度。

$$h_{mt,task} = W_{task}^T h_{mt} + b_{task} \tag{5-11}$$

$$c_{task} = W_{c,task}^T \sigma(h_{mt,task}) + b_{c,task}, \tag{5-12}$$

其中，$W_{task} \in \mathcal{R}^{d_t\times d_m}$ 为可学习参数，$b_{task} \in \mathcal{R}^{d_m}$ 为可学习的偏置参数，$W_{c,task} \in \mathcal{R}^{d_m}$ 为可学习中心度映射权重，$b_{c,task} \in \mathcal{R}$ 为可学习的中心度偏置，$task \in \{Sarcasm, Metaphor, Subtext\}$。

5.2.1.6 置信度预测层

在每个特征子空间中，预测每个 Token 的中心度，还需要对中心度是否可信进行评估。因此，本文在每个特征子空间中预测中心度的同时对其预测置信度 p。

$$p_{task} = W_{p,task}^T h_{mt,task} + b_{p,task}, \tag{5-13}$$

其中，$p_{task} \in \mathcal{R}$，$W_{p,task} \in \mathcal{R}^{d_m}$ 为可学习权重参数，$b_{p,task} \in \mathcal{R}$ 为可学习偏置参数。在预测时，设定阈值 δ，若 $p_{task} > \delta$，说明当前的预测中心度可信，保留其预测的结果；否则，当前预测的中心度不可信，预测的中心度置为 0。

5.2.1.7 损失函数

本文的损失函数，由两种不同的损失函数组成，其一为不同任务的中心度预测损失，我们将其损失设置为均方误差，记为 $\mathcal{L}_c$；其二为关于不同任务的置信度预测损失，我们将其当作分类任务进行，因此选用交叉熵 (CrossEntropy) 作为该部分的损失函数，记为 $\mathcal{L}_p$。如公式 5-16所示，最终的损失为两个损失

的加权。

$$\mathcal{L}_c = \sum_{task} \sum_{s \in \|S\|} \sum_{i=1}^{L_s} \frac{(c_i - \hat{c})^2}{2L_s \|S\|} \tag{5-14}$$

$$\begin{aligned} \mathcal{L}_p = & \sum_{task} \sum_{s \in \|S\|} -(\hat{p}_s \mathrm{I}(p_s > \delta) \log(p_s) \\ & + (1 - \hat{p}_s) \mathrm{I}(p_s <= \delta) \log(1 - p_s)) \end{aligned} \tag{5-15}$$

$$\mathcal{L} = \omega_c \mathcal{L}_c + \omega_p \mathcal{L}_p, \tag{5-16}$$

其中，S 为所有样本集合，s 为单个样本，$\|S\|$ 为样本的数量，L_s 为样本的内容长度。$\hat{c}$ 为真实的中心度值。$\hat{p}_s \in 0, 1$ 为真实的置信度值。$\mathrm{I}(x)$ 为指示函数，$\mathrm{I}(x) = 1$ 当且仅当 x 为真时成立。$\omega_c, \omega_p \in \mathcal{R}$ 为两个损失的权重。

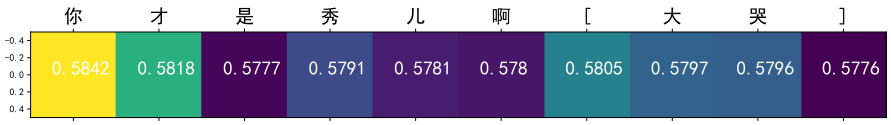
5.2.2 实验分析

本节旨在对潜台词片段识别的解决方案相关的实验做一个详细的介绍。主要分为以下几个部分：基线模型选取、评价指标的选取、实验细节、对比实验和消融实验。

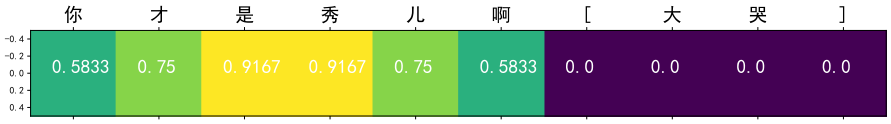
基线模型的选取

- **BERT+LSTM+Softmax+Multi-task**: 潜台词片段识别的单一任务可以简单的使用序列标注方法进行解决。考虑到我们任务和序列标注之间的关系，我们将序列标注模型中常用的算法使用多任务框架加以扩展，用做其中一个对比模型，简记为 BLSM。
- **基于边界的模型**: 潜台词片段识别和嵌套的命名实体识别任务具有结构上的统一性，因此引入边界相关的两个方法：Tan 等人[64] 提出的识别起止点的模型，记为 BENSC 模型；Zheng 等人[63] 提出的对每个 Token 识别 BIEO 中一种，利用多注意力机制和多任务结合的方法，记为 MHSA，作为我们的其中两个基线模型。
- **基于 Span 的方法**：嵌套类型的序列标注可以使用基于 Span 的方法进行识别，从问题的结构上而言，也可以用于对潜台词内容的识别。因此，引入当前最新的两段式方法 Locate-And-Label(LAL)[5] 作为我们其中一个对比方法。

评价指标的选取



(a) 中心度的预测结果



(b) 中心度的真实标注结果

图 5-5: 为什么需要引入 IoU 的样例

在类似于命名实体识别的任务中，评价指标通常为分类算法的评价指标 F_1。然而，潜台词内容和相关任务如命名实体识别的区别在于命名实体识别所需要识别的实体边界非常明确，而这恰巧不适用于潜台词的内容片段识别。如图 5-5所示的例子中，图 5-5(a)为模型所预测的结果，而图 5-5(b)为人工标注的结果，从语义层面而言，两者所对应的内容都可作为潜台词相关的子序列。但是在使用 F_1 进行评价时，该预测结果只能作为错误预测进行衡量，显然不合理。因此，本文增加并引入目标检测中常用的评价方式——Intersection-Over-Union(IOU)[81]，其计算方式如相关工作中所述，此处不再赘言。IoU 在目标检测中的作用是评价识别物体的准确性。因为物体具有一定的面积或者范围，如果要求每一个点位都匹配上才算识别正确，则太过于严格。IoU 则通过识别正确点位的占比来衡量识别的准确性，显得更为合理，而这同样适用于潜台词片段识别。除此之外，本文基于 IoU 重新计算了 F_1 指标，给定一个阈值 θ，当 $IoU(x, \hat{x}) > \theta$ 时，x 被判定为正确的正样本；当 $0 < IoU(x, \hat{x}) < \theta$ 时，x 为错误的正样本；当 $IoU(x, \hat{x}) = 0$, $IoU(x, x) = 1$ 且真实标注结果无正例时，为错误的负样本；当 $IoU(x, \hat{x}) = 0$, $IoU(\hat{x}, \hat{x}) = 1$ 且预测结果无正样本时为错误的正样本。

实验细节

本章所用数据集源自于本文第三章所述工作，从中抽取出潜台词、比喻和反讽相关的样例进行训练。为了减缓模型过拟合的数量，从其余无关的数据中随机选取 1/10 的负样本，总计数据量为 3270 条，并按照 8：2 的比例，划分为训练集和测试集。为了减缓模型过拟合的问题，CBM 对需要分析的文本内容 W 和完整的输入序列 S 都采用了 dropout 技术，并且赋予了不同的比例，对于完整输入序列，由于其序列较长，且目标内容占比较少，因此我们赋予较高的

dropout 比例。除此之外，CBM 还使用了 Early Stop 的方法进一步减缓过拟合问题，最终的呈现结果为在验证集在 Early Stop 条件下衡量结果最优所对应的模型，在测试集上的衡量结果。对于衡量的方法，我们结合使用分类的衡量指标 F_1 和 IoU 进行评估。在我们的实验中，我们使用单张 GPU(GTX 1080Ti) 进行训练，优化方式为 AdamW，实验中涉及的超参数，如表 5-1所示。

表 5-1: 超参数表

Embedding Size		GloVe	300	BERT	768	Positional	50
Train:Validation:Test		6:2:2		k	3	δ	0.5
learning rate		1e-2		warm up step	50	epoch	50
Random seed	2021	batch size	8	ω_c		0.5	
dropout rate for W	0.2	dropout rate for S	0.4	ω_p		0.5	
cross validation fold		-		patience for early stop		10	

5.2.2.1 对比实验

本节将 CBM 模型和经典的平展类型序列标注方法和当前最新的嵌套结构的实体识别算法进行对比。为了比较的全面性和公平性，选取了一种经典的平展类型序列标注算法（BLSM）和三种嵌套类型的标注算法（BENSC、MHSA、LAL），每种算法代表不同的建模思路，且在相应任务上都取得了不错的效果。对比实验的结果，如表 5-3所示。对于有开源代码的模型，我们将数据处理成相应格式，并且去除本文无法提供的输入项，以适用于我们的任务；而对于没有开源代码的方法，我们对比其论文中所述进行实现。表 5-2和 5-3中，所有带有 * 的均为本文自己实现并且对参数进行二次调优后的结果。带有下划线的为本文所提的方法结果。除了衡量一些常见指标之外，我们还衡量了在同一个测试集上的推理时间，展示为表中的 Inference Time 一列。

表 5-2: 不同模型在潜台词片段识别上的 IoU 评估结果

Model	IoU_{subt}	IoU_{sarc}	IoU_{meta}	Inference Time
CBM	0.5690	0.4897	0.3467	40 s
BLSM	0.5222	0.4421	0.3088	37 s
BLSM*	0.5325	0.4482	0.3012	38 s
BENSC*	0.5198	0.4311	0.3010	41 s
MHSA	0.5198	0.4524	0.3175	65 s
MHSA*	0.5245	0.4535	0.3117	75 s
LAL	0.5485	0.4551	0.3378	125 s

表 5-2所展示的是不同模型的预测结果在 IoU 上的计算结果。其中除 CBM 外，其余模型由于都是对 Span 或者 Token 进行的分类操作，因此，可以认定其中心度为非 0 即 1，则可以与 CBM 在后续计算上达到统一。从 IoU 的统计结果中可得，所有模型关于普通类型的潜台词识别准确率，仅有 0.5 左右，这显然没有达到预期的目标。我们对 CBM 的模型进行分析，发现其尽可能长的预测潜台词、反讽或者比喻相关的内容。这导致了尽管预测的内容可能覆盖到整个真实结果，但是由于预测的内容较长，而使得整体的 IoU 偏低。对于其余模型，则产生的原因不一，其中预测结果有部分为仅与真实结果相交，另一部分为预测结果被真实标注结果所覆盖，即预测范围长度不足。另外所有模型的共同不足之处在于，都有一部分的内容没有识别出来。

表 5-3: 不同模型在潜台词片段识别上的 F_1 评估结果

Model	$p_{subt}(\%)$	$r_{subt}(\%)$	$F_{1,subt}(\%)$	$p_{meta}(\%)$	$r_{meta}(\%)$	$F_{1,meta}(\%)$	$p_{sarc}(\%)$	$r_{sarc}(\%)$	$F_{1,sarc}(\%)$	$Macro-F_1$**(%)**
CBM	73.44	41.59	53.11	56.83	16.36	25.40	87.83	15.81	26.79	36.71
BLSM	68.13	39.49	50.00	58.27	13.80	22.37	84.35	15.25	25.83	33.84
BLSM*	68.75	39.71	50.34	58.99	13.80	22.37	84.35	15.25	25.83	33.84
BENSC*	70.75	38.71	50.04	58.27	13.59	22.04	84.35	15.25	25.83	33.73
MHSA	69.53	39.71	50.55	59.71	13.97	22.65	84.93	15.55	26.28	33.93
MHSA*	68.85	39.78	50.43	61.15	14.29	23.16	84.93	15.55	26.28	34.08
LAL	70.95	40.62	51.66	61.15	14.39	23.30	84.38	15.25	25.83	34.34

表 5-3中展示了所有的对比模型的 F_1 评估结果，均为基于 IoU 所得。$Macro-F_1$ 为模型在各个任务上的总体衡量分数。从表中数据可知，所有的模型对于识别普通类型的潜台词都能够做的比比喻和反讽类型的潜台词更好，其中 CBM 具有最好的效果，LAL 次之。比喻和反讽类型的潜台词识别，每个模型都不能得到很好的结果。同时，从表中可得，大部分模型都存在的问题为模型识别的准确度较高，但是召回率却远不能满足需求。类似于 LAL 这类的模型直接判断获得的子序列是否为目标序列，再更具 IoU 做最终判定，使得其以更小的概率得到和真实标注接近的 Span。因此，这类算法在 IoU 和 F_1 上的效果都不如 CBM。而 CBM 不需要预先判断子序列，而是直接根据 IoU 判定当前内容的子序列是否为正确的样本。除此之外，我们采用中心度的理论下限作为过滤的阈值，也使得判定的序列长度可以在预测之后，对预测序列的边界进行微调，而不至于类似于基于 Span 的判断方式，以 Span 为单位进行判定。

5.2.2.2 消融实验

本节探究本文所提出的中心度和深度特征融合层的效果。本节进行三个消融实验，一为我们将中心度替换成二分类，即判断某个 Token 是否属于某个特殊内容，记为 BCBM；二是将中心度替换成 BIEO 的标签预测，记为 SLBM；三是不使用 DFML 层，而是使用上下文特征拼接，通过 MLP 降维的方式进行替换，记为 CMBM。消融实验的 IoU 评估结果如表 5-4所示，F_1 评估结果如表 5-5所示。

表 5-4: 消融实验在 IoU 上的评估结果

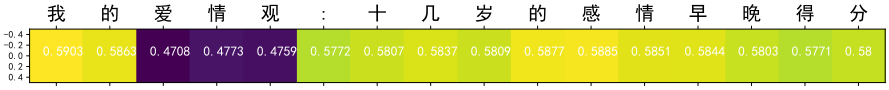
Model	IoU_{subt}	IoU_{sarc}	IoU_{meta}
CBM	0.5690	0.4897	0.3467
BCBM	0.5479	0.4773	0.3246
SLBM	0.5532	0.4690	0.3306
CMBM	0.5554	0.4725	0.3240

我们将回归任务变成非 0 即 1 的分类任务后，从表 5-4中可得，其效果略低于使用中心度预测的结果。通过对样例的识别结果进行分析，如图 5-6中所示，BCBM 预测的结果和真实的结果（“十几岁的感情早晚得分”）有部分为存在交集，但不完全覆盖真实标注的结果；而 CBM 所预测的结果，大多完整包含真实结果，但会比真实结果所覆盖的范围会更广。

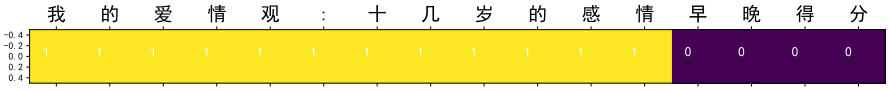
表 5-5: 消融实验在 F_1 上的评估结果

Model	$p_{subt}(\%)$	$r_{subt}(\%)$	$F_{1,subt}(\%)$	$p_{meta}(\%)$	$r_{meta}(\%)$	$F_{1,meta}(\%)$	$p_{sarc}(\%)$	$r_{sarc}(\%)$	$F_{1,sarc}(\%)$	$Macro-F_1$**(%)**
CBM	73.44	41.59	53.11	56.83	16.36	25.40	87.83	15.81	26.79	36.71
BCBM	70.63	40.57	51.54	61.15	14.36	23.26	85.22	15.48	26.20	34.72
SLBM	71.25	40.71	51.82	60.43	14.21	23.01	85.22	15.43	26.13	34.69
CMBM	71.88	40.78	52.04	56.12	13.81	22.16	86.09	15.54	26.33	35.29

基于 BIEO 标注法进行设计的 SLBM 模型，在这方面略好于 BCBM，可见当任务细化之后，是有助于提升模型的效果的，但是由于数据量等原因，提升不是很明显。从表 5-4和 5-5中均可得，深度特征融合方式相比于传统的双向 LSTM 特征融合方式在三个任务上都有略微有所提升，说明深度特征融合层在对上下文建模信息的特征融合与抽取上，比传统的特征融合方式更好。



(a) CBM 的预测结果



(b) BCBM 的预测结果

图 5-6: CBM 和 BCBM 的预测样例展示

5.2.2.3 原因分析

从对比试验和消融实验中，我们都可以得到，不论是 CBM 或是其余对比模型，在反讽和比喻上都不能做得很好。本文对相关样例进行分析，得到以下的几个原因：

数据方面: 1. 本节中采用的数据集，源自于第三章所述内容。从数据进行分析可知本文可用的数据量是较小的，对于一个与语义高度相关的任务而言，这个数据量显然不够。因此，本文也尝试了使用数据增强等方法对样本进行扩容。考虑到潜台词的语义可能被其中任何一个词影响，因此，数据增强的部分仅对一些标点符号进行替换。我们将增强后的数据进行实验得到 $Macro-F_1$ 数值为 35.73，$F_{1,sarc}$ 为 28.31，$F_{1,meta}$ 为 23.32，$F_{1,subt}$ 为 52.04。尽管潜台词识别的效果略有下降，但是总体略有提升。可见若有更好的数据增强手段和有更多的数据量，识别的效果还可以继续提升。2. 由于反讽和比喻占比较少，在模型训练过程中，大多见到的为不含有反讽和比喻的样例，因此导致模型学习到的结果有偏；尽管我们在模型训练的过程中采用了代价敏感分析等手段，但依旧不能让它很好的识别出来。而由于普通的潜台词在我们构建数据集时占比较多，而不存在这类问题，因此潜台词分类的效果要好于其余两个任务。

模型原因：由于反讽和比喻占比较少，为了防止模型学习到的结果有偏，除了使用代价敏感分析以外，我们还对其进行了两种尝试，其一为当一批数据中不具有比喻或者反讽时，则当前批样本不产生对于反讽和比喻的分类损失；而当句子中含有反讽和比喻时，我们对负样本按照 3：1 的负正样本比进行采样，使之稍微平衡。其实验结果便为本文所示的结果。然而，这种方法产生的问题为在测试时，尽管一个样本不含有比喻或者反讽，其依旧有时会有一个较高的置信度预测其中有比喻和反讽的部分。其二为当一批数据中不具有比喻或者反讽

时，我们按照一定概率选取其中其中某个样本使之产生损失。这个概率难以确定，本文尝试了 5%、10% 与 20% 的概率，但都没有很好的结果。发现，当概率较小时，其和不对该类样本进行采样效果相当；而当概率较大时，则引入了过多的负样本。另外，这种方法在训练过程中，也会使得网络的损失值不能稳定下降，引起震荡。

5.3 基于中心度的嵌套命名实体识别

嵌套命名实体识别的任务形式，于潜台词片段识别相近，区别之处在于，实体有着明确边界，相比于潜台词片段识别其几乎没有模糊性。关于嵌套命名实体识别这一个任务，有很多的工作被提出（如相关工作中所述）。尽管都取得了不错的效果，然而却还有其中缺失和不足之处。其缺失之处为大部分的方案都没有考虑到实体与实体之间的联系。基于层级结构的方案可以自然地建模从长度较短的实体向长度较长的实体地转变，如 Pyramid 模型[4] 和 PO-TreeCRFs 模型[67] 等，但是这些模型都没有进一步对其进行建模，仅让模型自主地从短往长的方向进行学习。这是比较低效而且非常隐晦的方案。其不足之处如前所述，无论是基于层级结构的方案或者是基于枚举 Span 的方案，都面临着一个问题，即消耗大量的时间和空间。因此，基于这两点，本文提出将中心度的方法用于嵌套命名实体识别，并且显式地建模长短实体之间的联系。本文称该部分算法为“带有中心度的层级模型 (Centerness Based Layer Model for Nested Named Entity Recognition, CBLM)”用于嵌套命名实体识别，模型的结构如图5-7所示。此外，与上一节所述不同，本节所采用的模型并非于多任务框架，而是所有的可能实体都掺杂一处，因此对于 centerness 的定义也应是一个综合体现的结果。为此，对其做一个简要的修改，设 $E = \{e_{i,1}, e_{i,2}, \cdots, e_{i,k}\}$ 为某个 Token i 所在的 k 个实体，每个实体所对应的中心度分别为 $c_{i,1}, c_{i,2}, \cdots, c_{i,k}$，则对于 Token i 而言，其中心度值越小，则其越靠近边界；反之，其越靠近实体中心。如果要减少枚举 Span 产生的时间和空间消耗，则更应该关注当前区域内的 Token，是否可能为某个实体的一部分，因此，选取综合的中心度为 $c_i = \max(c_{i,1}, c_{i,2}, \cdots, c_{i,k})$ 为 Token i 的中心度表示，可以让其与不重要的 Token 差别最大化，利于模型学习。

5.3.1 模型结构

本节主要介绍 CBLM 的模型结构，旨在说明如何在层级结构中使用中心度并且说明如何建模长短实体之间的联系。模型主要分为几个部分：编码层、上下文建模层、过滤层、跨层注意力模块和预测层。编码层本文采用和 LAL 模型[5] 一致的设定，其中本文不涉及提前获取 Span 表征的操作。上下文建模层采用双向 LSTM 进行建模，与章节 4中所述一致，不再赘述。其余模块详见后文论述。

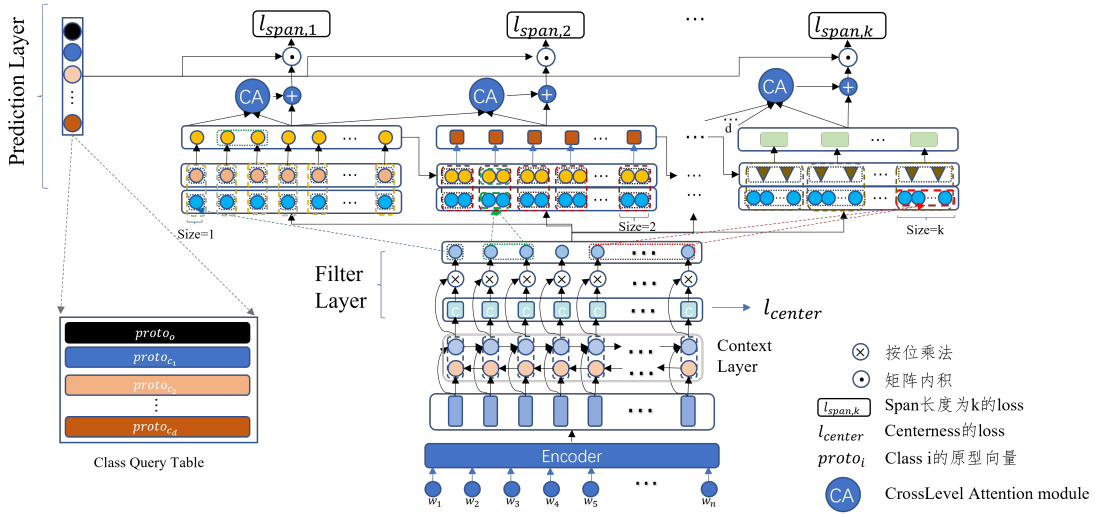


图 5-7: CBLM 模型结构图

5.3.1.1 过滤层

设经过双向 LSTM 处理后得到的语义特征向量为 h，词性特征向量为 h_{pos}，对每个位置的特征 h_i 和每个位置的词性特征 $h_{i,pos}$ 进行拼接，通过 MLP 降到一维空间之中得到每个位置对应的中心度 c_i。然后将中心度作为过滤的条件，与原来的特征相乘，得到带有中心度的特征，向层级结构中的上一层进行传递，其计算如式 5-17到 5-18所示。

$$c_i = MLP([h_i, h_{i,pos}]) \tag{5-17}$$

$$h_{c,i} = c_i \times h_i. \tag{5-18}$$

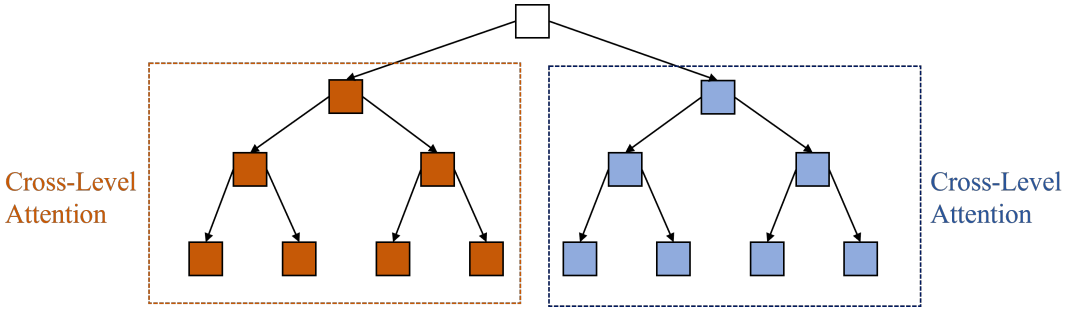


图 5-8: 跨层注意力机制图示

在测试阶段，设定一个阈值 δ，当预测的 c_i 低于 δ 时，删除当前位置的特征，后续枚举过程不再考虑当前位置相关的所有可能。以此达到过滤的效果。

5.3.1.2 跨层注意力层

跨层注意力层 (Cross Level Attention Layer) 的目的是去计算相互有关联的 Span 之间的关注度，使得当在对长实体进行判断时，可以更有效的利用与其相关的短实体信息。另外，本文假设，当短实体和长实体之间相差过长时，可认为两个实体之间不存在必然联系。因此，本小节所述的跨层注意力仅关注 k 层以内的实体转换关系。在层数相差 k 层以内，跨层注意力层的计算如图 5-8所示，其中越处于底层，实体长度越短，越高层实体长度越长；褐色部分和蓝色部分的注意力分别以当前子树的顶点作为查询 (Query)，以子树的内部节点和叶节点作为键 (Key) 和值 (Value)。从图的角度而言，图 5-8是一个有向图，跨层注意力层计算了从某个节点出发，k 步可达的节点之间的注意力关系，是图注意力机制的一种。

计算跨层的注意力，需要先得到每层 Span 的特征表示。第 i 层表征由三部分构成，其一为前一层表征中通过长度为 2 的窗口加权求和得到；其二为通过带有注意力的表征 $h_{c,i}$ 进行最大池化得到；其三为由原始序列所对应的词性标签序列 h_{pos} 得到。第 i 层第 j 个位置的表征计算如式 5-19所示：

$$r_{i,j} = [r'_{i-1,j}; r'_{i,j}; r'_{i,j,pos}] \tag{5-19}$$

$$r'_{i-1,j} = W^T[r_{i-1,j}; r_{i-1,j+1}] \tag{5-20}$$

$$r'_{i,j} = MaxPooling([h_{c,i,j}; \cdots; h_{c,i,j+i}]) \tag{5-21}$$

$$r'_{i,j,pos} = MaxPooling([h_{j,pos}; \cdots; h_{j+1,pos}]). \tag{5-22}$$

其中 $W \in \mathcal{R}^{2\times d}$ 为可学习权重，d 为特征表示的维度。得到相应层的特征表示

算法 5.1 掩码生成

输入： k : 可见的最大层数; s : 句子的最大长度
输出： masks

```
tl ← k * (2s+1+k)/2
masks ← torch.ones((1, s, tl))
i ← 0; j ← 1
repeat
  base ← 0
  repeat
    masks[i, i + base : min(s + base + j, i + base + j + 1)] ← 0
    base ← base + s + j
    j ← j + 1
  until j = k + 1
  i ← i + 1
until i = s
```

后，便可对不同长度实体进行转化关系使用 Attention 的方式进行建模。由于不同层相连接的 Span 不具有特殊结构可以直接计算，因此本文使用掩码的方式实现关联关系的构建。构建的方法如算法 5.1所示。获取了相应的 mask 之后，利用 Self-Attention 的计算方式对其进行建模。其过程如式 5-23所示。

$$r_{i,span} = att^T v \tag{5-23}$$

$$att = masks \times Softmax(\frac{q^T k}{\sqrt{d}}) \tag{5-24}$$

$$q = W_q^T r_i,\ k = W_k^T r_i,\ v = W_v^T r_i, \tag{5-25}$$

其中 $w_q, W_k, W_v \in \mathcal{R}^{d\times d}$ 为可学习的权重参数，$\times$ 为按位乘法。最后，通过残差连接的方式和上一层的表征进行 r 相加得到最终每一层的表征。

5.3.1.3 预测层

不用以往分类算法，本节不直接对每一层得到的特征进行分类计算，而是先随机初始化一组类别相关的原型向量，如图 5-7所示。通过计算每一层特征和原型向量的相似度，选取最大的为最后的分类结果。设原型向量为 $Proto$，则分类计算方法如式 5-26所示。

$$y_i = \text{argmax}\, Softmax(Proto^T(r_{i,span} + r_i)). \tag{5-26}$$

5.3.1.4 损失函数

CBLM 的损失函数，有两部分构成，其一为中心度的预测损失；其二为针对每个 Span 的分类损失。其中，中心度的预测损失采用均方误差（Mean Square Error，MSE）, 分类损失采用交叉熵损失（Cross Entropy Loss, CEL）。

5.3.2 实验分析

本文使用的数据集为 GENIA，该数据集为开源的生物领域的嵌套命名实体识别的数据集，包括了五类实体类别：DNA、RNA、protein、cell_line 和 cell_type。本文将该数据集按照 Yu 等人[82] 的设置进行划分，划分比例为训练集占 90%，测试集占 10%。本实验的结果均为 P100 GPU 上运行所得。训练和测试时设置的超参数如表 5-6所示。

表 5-6: CBLM 的超参数表

Epoch	80	max length	512	batch size	24
learning rate	3e-5	random seed	42	centerness threshold	0.35

5.3.2.1 对比实验

本文将 CBLM 模型和嵌套命名实体中当前最新的模型进行比较，其中包括基于边界预测的模型 MHSA[63] 和 BENSC[64]、基于层级结构的模型 Pyramid[4] 和 PO-TreeCRFs[67] 和基于两段式的模型 LAL[5]。结果如表 5-7所示，其中 Speed 一栏描述的是模型的推理速度，其单位为每秒执行的词的数量 (words-per-second, w/s)，该项数值越大越好。从表 5-7中结果可得，本文所提 CBLM 模型和 LAL 模型在嵌套的命名实体上效果相近，远超其余模型。虽然 LAL 模型的效果略好于我们的模型，但是从运行速度上，我们的模型却远好于其。

关于 CBLM 和 LAL 的差别，本文对两者之间所不能解决的样例进行了统计分析，发现 CBLM 中所擅长解决的问题为长实体的识别，而对于长度非常短的实体识别效果略差于 LAL。

5.3.2.2 消融实验

为了验证本文所提方案所带来的有效性，本文还设置了以下几组消融实验，分别为仅有层级结构且直接分类的预测模型（BASE）、带有跨层注意力的模型（BASE+CLA）、使用原型向量的方法预测的模型（BASE+CLA+Q）、使用分类作为过滤层的跨层注意力模型（BASE+CLA+F）、使用中心度作为过滤层的跨层注意力模型（BASE+CLA+C）和 CBLM 模型。其实验结果如表 5-8所示。 从表中可得，本文所提出的跨层注意力层不引入其余特征即可取

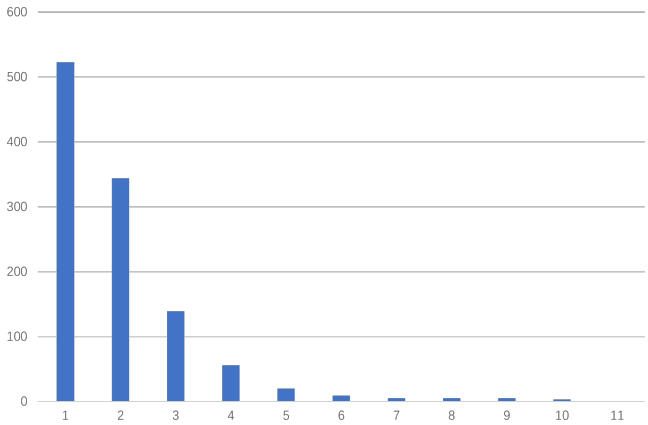


图 5-9: 错误样例的长度统计结果。其中横轴代表错误样例的长度，纵轴代表该长度的错误样例数目。

得 0.93% 的效果提升，引入原型向量作为分类预测的方式可以带来 0.15% 的效果提升，引入中心度的方法可以带来 0.56% 的效果提升，完整的模型，在这基础之上每一层增加词性特征，可以带来 0.47% 的效果提升，总提升为 2.08%。消融实验 BASE+CLA+F 说明在模型运行过程中采用非 0 即 1 的过滤方式有助于提升模型的效果。消融实验 BASE+CLA+C 在 BASE+CLA+F 对比后更可以说明过滤有助于模型更好的判别，同时软性的过滤可以更好的保留特征，使之有更好的效果。完整的模型 CBLM 和其余消融实验对比，也说明更丰富的特征和

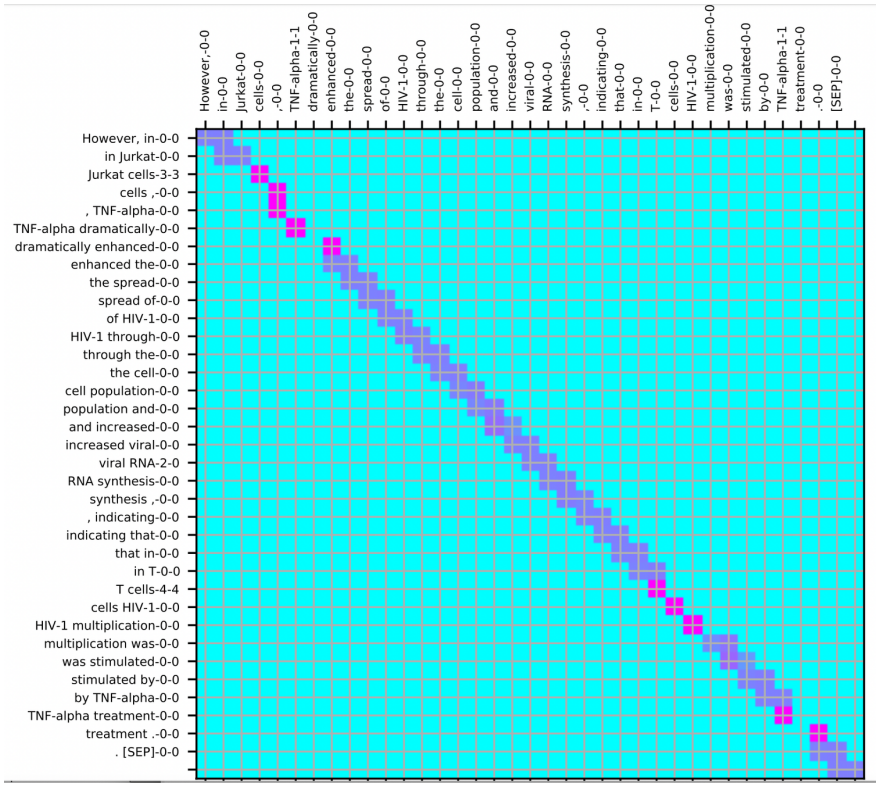
表 5-7: 嵌套命名实体识别的对比实验结果

Model	Precision(%)	Recall(%)	F_1(%)	Speed
Pyramid	80.3	78.3	79.3	3280 w/s
LAL	80.2	80.9	80.5	2582 w/s
BENSC	79.2	77.4	78.3	10675 w/s
PO-TreeCRFs	78.2	78.2	78.2	1897 w/s
MHSA-BERT	80.3	78.9	79.6	10798 w/s
Ours	82.1	78.6	80.3	10547 w/s

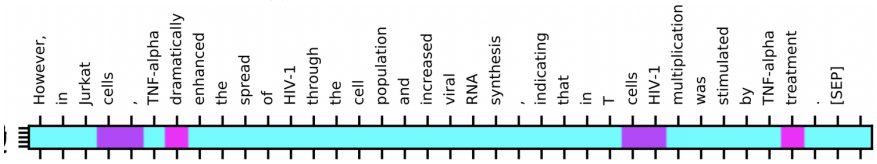
表 5-8: CBLM 的消融实验结果

Model	Precision(%)	Recall(%)	F_1(%)
BASE	79.41	77.75	78.20
BASE+CLA	81.10	77.24	79.13
BASE+CLA+Q	80.83	77.79	79.28
BASE+CLA+F	81.42	77.42	79.37
BASE+CLA+C	81.38	78.35	79.84
Ours	82.15	78.55	80.31

更有效的特征对于提升模型效果有着至关重要的作用。



(a) 跨层注意力层得到的注意力值



(b) 过滤层产生的中心度预测结果

图 5-10: 跨层注意力值和中心度

本文对于 CBLM 不能取得更好效果的原因做了进一步的分析，发现主要有以下两个原因。其一为 CBLM 擅长识别较长的实体，这是由 CBLM 的结构所决定的，长度越长其可以得到的特征信息就越多，因此其判断起来就更加容

易。但是相对而言，CBLM 在处理长度较短的实体就稍显不足。图 5-9所示为 CBLM 识别错误的样例统计信息，横轴代表样例的长度，纵轴代表该长度下样例出错的数目，经统计，最长实体长度为 30，图中仅展示了识别错误的样例的统计信息。从图中可得，大部分情况下，CBLM 犯错发生在长度为 3 以内样例中，而对更长的样例则更少犯错甚至几乎不会犯错。

其二为中心度在带来很好的效果时，其也有一定的不足。图 5-10所示为 CBLM 所识别错误的一个样例可视化信息，图 5-10(a)所示为跨层注意层计算所得的注意力值，图 5-10(b)所示为过滤层所得的中心度信息，可见注意力关注信息是正确的，但是其所计算的中心度略有偏差，倒是相应的实体被中心度所过滤，而没有被识别出来。经统计，出于该部分原因而识别错误的占比达到了 52.96%。

5.4 本章小结

在本章中，我们通过详细的例子具体说明了什么是潜台词片段识别任务和什么是嵌套命名实体识别任务，并对该类任务进行抽象化，提取了共同的特征。根据相应方案的不足，我们提出中心度的概念。并将之应用在潜台词片段识别任务和嵌套命名实体识别任务上，并且根据两者的特点，分别设计了两个解决方案。其中关于潜台词片段识别部分，所有的相关模型都不能取得很好的效果，本文对目前解决方案的不足进行了分析，并且提出了未来工作的研究方向。中心度在嵌套命名实体识别上的效果很好，在相同的运算效率下，CBLM 取得了最好的效果，在相同的效果中，CBLM 具有最快的运算效率。

第六章　数据搜集与潜台词分析系统

为了验证本文所提出算法在实际推荐场景中的有效性，我们搭建了一个潜台词识别演示系统并将本文所提出方法应用其中。下文将对潜台词识别演示系统的背景、功能、整体架构、实现细节及识别效果等进行详细介绍。

6.1　相关背景

在智能设备日益普及和智能化程度越来越高的时代背景下，人们对于聊天机器人的需求在增加，对于聊天机器人的智能化的要求也在提高。因此，让聊天机器人读懂人类的潜台词具有深远的意义。出于本文数据和当前的研究技术的考虑，本文仅提供聊天机器人的前期功能——即识别出人类对话中是否具有潜台词以及哪个部分具有潜台词。这个前期功能是聊天机器人能否正确理解人类潜台词的基础。当聊天机器人具有判别当前语句中是否具有潜台词的能力时，尽管不知道具体是什么内容，但是在对应答内容进行挑选的时候，就可以将是否具有潜台词作为一个考虑的维度，进而过滤一些低质量的回答。更有甚者，可以使用一定方法，让聊天机器人的回答根据潜台词的类型而不同，使得聊天的内容更具有趣味性。

真实的聊天系统在使用时都需要有发起方，要么是使用者主动发起，要么是聊天系统根据预设内容开启对话。但无论哪种情况，在开始阶段都是没有上下文可以供给参考的，无法进行有效的潜台词分析；而当聊天轮次增加时，就可以构建出一个相对完整的上下文信息和背景信息，此时对潜台词的分析才相对准确和有意义。本文用于研究的场景和上述所说场景不尽相同，区别在于本文假设当前分析内容的上下文信息如背景信息、场景信息或者上下句信息等都具备，不需要经历对话积累的过程，对应到真实场景中中后期阶段。同时，为了后续研究方便，本文还提供数据搜集的功能，方便使用者进行上传数据和下载数据用于研究所用。

6.2 系统设计

本文结合本文研究的发展过程，搭建了一个可以进行数据搜集、数据下载和潜台词检测以及潜台词片段识别的潜台词分析演示系统，并将本文第四章与第五章中与潜台词相关的算法应用其中，实现了输入一句话及其相关的上下文信息，即可自动化分析该句是否具有潜台词、反讽和比喻的目的，同时还可以自动化分析出潜台词、反讽和比喻所对应的内容。

6.2.1 系统需求

潜台词分析演示系统的功能包括可以简要归结如下：

- 数据录入：所有使用本系统的用户都可以贡献自己的数据。为了避免不同用户之间习惯用于的不同，导致数据上的差异，则应让用户在使用时，可以直接从制定的标准中进行选取。
- 数据下载：所有用户录入数据之后，需要有一个集中呈现的页面，使用户可以进行浏览和下载。
- 潜台词分析功能：用户可以使用本系统进行自动化分析语句，同时给用户呈现的分析结果应该直观且显眼。

6.3 系统实现与效果展示

潜台词分析演示系统实现分为前端和后端两个部分。前端基于开源的 React 框架以及 Ant Design 组件库进行实现，后端使用基于 python 的轻量级后端框架 FLASK 进行实现。前端的编程语言为 HTML 和 javascript，后端的编程语言为 python。除此之外，本文算法部分分别使用开源的深度学习框架 Keras 和 PyTorch 进行实现。下文将根据三个系统需求分别展示我们的系统。

6.3.1 数据录入模块

数据录入功能，需要用户提供两方面的原始信息，其一为可用于分析的目标语句，其二为该目标语句所处的完整上下文信息。另外，由于该系统是为了本文研究所服务，因此按照本文的数据要求，需要用户提供目标语句的是否含有潜台词、比喻、反讽、夸张、谐音、情绪和态度等类别信息。该部分，本系

统采用下拉框的形式实现，减轻用户的负担和避免用户在录入情绪词时产生的描述不一致现象。图 6-1展示的为该部分的前端效果。用户在录入所有信息之后，可以点击提交进行上传数据；或者点击取消重新进行录入。上传的数据采用 Mysql 数据库进行存储。

图 6-1: 数据录入模块效果展示

6.3.2 数据展示模块

数据展示功能可以让用户快速的获得所有已有数据，并且可以根据不同的类别标签进行快速筛选。数据的结构化形式较好，本系统使用表格的形式对数据进行展示，同时提供每列排序的功能。用户可以通过点击每列的表头按照升序或者降序进行排列。除此以外，用户在新录入数据后，需要通过刷新操作，获取当前最新的数据。当有需要时，用户可以点击下载，将本系统中的所有数据以 *Json* 的格式保存到本地。图 6-2展示的为该部分的前端效果。

图 6-2: 数据展示模块效果展示

图 6-3: 潜台词分析模块的效果展示

6.3.3 潜台词分析模块

潜台词分析模块提供两个算法的分析效果，其一为判断目标语句是否为潜台词、比喻或者反讽；其二为识别其相关的原文内容。分析的目标语句与数据录入中所示一致，为一个文本输入框；目标语句的上下文信息同样也是用一个文本输入框进行获取。输入内容后，用户可点击提交预测，将数据发送给后台进行分析处理；也可以点击取消，进行重新录入。后端算法侧将分析结果返回前端，前端做相应的可视化处理。前端于该部分展示目标语句是否为潜台词、比喻和反讽，并且使用三个百分比可视化图展示后台预测的概率情况。同时，前端使用不同的颜色在原文上标注出其是否是属于潜台词的内容或是否属于比喻和反讽的内容。图6-3展示的是潜台词分析模块的前端效果。图6-4展示的潜台词分析结果的详情，其中最左边为潜台词检测的识别结果，展示了判别的结果和其对应的概率。中间部分为潜台词片段识别的结果，使用于标签相同的颜色标注出了原文中哪部分是潜台词、比喻和反讽。最右边为根据识别结果，所推荐的问句回复。

图 6-4: 潜台词分析的结果展示

6.4 本章小结

本章介绍了本文所搭建的潜台词分析演示系统。该系统的主要功能有数据录入、数据下载、潜台词检测分析和潜台词片段识别分析。该系统集成了本文所提的两个任务及其相应算法：潜台词检测任务与 SASICM 算法和潜台词片段识别与 CBM 算法。实践效果表明，本文所提算法可以提供良好的潜台词检测的能力、普通潜台词片段识别的能力，在比喻和反讽内容上的识别能力稍显不足，还有改良空间。在当前条件下，该系统依旧表明了本文所提算法的应用价值。

第七章　总结与展望

潜台词在日常生活中使用非常广泛，它具有语义隐晦、特征难以提取等特点。调研结果显示，在自然语言处理领域中，本文是首次对其展开研究的工作。因潜台词在日常交流和文学作品中使用广泛，且在不同的语言中都有类似的现象，本文认为对潜台词的研究是必要的，因此提出了潜台词分析的系列问题，其中包括潜台词检测、潜台词片段识别和潜台词还原。为解决潜台词分析的系列问题，本文构建了中文潜台词数据集，命名为 CSD 数据集，其可以用于多个任务如：比喻分析、反讽分析、情感分析、情绪分析等。基于 CSD-数据集，本文对潜台词检测、潜台词片段识别进行了探究。关于潜台词检测，本文基于所提出的强化注意力机制 (Strengthen Attention) 设计了对目标语句和上下文语句分别建模的模型 SASICM。SASICM 在潜台词检测任务上取得了良好的效果，并且本文还对 SASICM 和其余对比模型之间进行了详细的分析，阐述了 SASICM 取得较好效果的原因。更进一步，本文提出了潜台词片段识别任务，并且对潜台词片段识别任务和嵌套命名实体识别任务进行了分析，分析得到共性，并根据该共性提出了中心度的概念。在文中，我们基于中心度根据潜台词片段识别任务和嵌套命名实体识别任务的各自特点，分别提出了相应的解决方案，其中“基于中心度的多任务片段识别模型 (CBM)”用于解决潜台词片段识别问题；“基于中心度的层级模型 (CBLM)”用于解决嵌套命名实体识别任务。对于潜台词片段识别任务，CBM 取得了比其余模型都要更好的效果，但是还是有其不足和解决不好之处，本文在相应章节探究了问题所在。在嵌套命名实体识别任务中，CBLM 模型兼顾了效率和正确率，做到了速度相当时，效果最好；效果相近时，速度最优，并且不需要复杂的预处理。同时，文中也分析了 CBLM 的不足之处。最后，本文将潜台词检测任务和潜台词片段识别任务及其相应模型统一集成到了演示系统之中。该系统提供了数据搜集、数据下载和潜台词分析等功能。

按照本文的现有进展，我们根据各任务总结分析了以下可以继续提升改进之处：

- 数据集部分：本文所收集的数据均来源于社交媒体，对于一些正式化的文

学作品却无相关的数据。因此，为了方便将来研究者的使用，可以扩充其数据类型至文学作品中。其次，本文经过筛选之后的数据，从数据的标签分布可知，有较多类别的数据量较小，因此后续可以针对性的扩充相应的类别数据。

- 潜台词检测部分：本文提出了 Strengthen Attention 的概念，但是从其 Attention 的分布可知，虽然其可以更有效的区分不同词之间的重要性信息，但是其和 Self-Attention 所共有的问题在于，大部分词都只是关注到自己，而对其余词关注较少。因此，在对 Strengthen Attention 的建模上，后续还有继续优化的方向，其中包括但不限于如何让注意力的分布更加合理、是否有其余建模注意力的方式等。其次，本文对于隐含语义的建模通过采用非线性映射的方法进行实现，在同一个语义空间中进行特征表示。这虽然有效，但是却难以解释其具体的语义，因此在这方面可以继续进行探究，探究是否可以将语义空间拆解成几个可解释的维度，并且分别进行建模。
- 基于中心度的片段识别：本文基于中心度提出了两个算法分别解决了两个问题。然而，中心度虽然带来了效果上的提升，但还有其不足之处。其一为中心度的中心可以定义的更加合理。本文定义中心度时直接采用其序列中心作为中心点，对于语义理解而言，这显然不是一个最好的方式。更好的方式应该是找到其语义中心，从语义中心出发，进行建模。此为第一个可以探究的点。其次，中心度随着原理中心点而降低，且是呈现线性降低的效果，这是否合理，抑或是非线性方式进行降低才是更好的方式。此为第二个可以深入探究之处。另外，对于潜台词片段识别，其效果没有预想中的好，且对比模型也没有取得好的效果。对于如何更好的判别潜台词的识别内容将是一个长久探究课题。

参考文献

[1] PLUTCHIK R. Emotions: A general psychoevolutionary theory[J]. Approaches to emotion, 1984, 1984 : 197 – 219.

[2] NAKOV P, ROSENTHAL S, KOZAREVA Z, et al. SemEval-2013 Task 2: Sentiment Analysis in Twitter[C] // Proceedings of the International Workshop on Semantic Evaluation. 2013 : 312 – 320.

[3] XU Y, HUANG H, FENG C, et al. A Supervised Multi-Head Self-Attention Network for Nested Named Entity Recognition[C] // Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. [S.l.] : AAAI Press, 2021 : 14185 – 14193.

[4] WANG J, SHOU L, CHEN K, et al. Pyramid: A Layered Model for Nested Named Entity Recognition[C] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online : Association for Computational Linguistics, 2020 : 5918 – 5928.

[5] SHEN Y, MA X, TAN Z, et al. Locate and Label: A Two-stage Identifier for Nested Named Entity Recognition[C] // ZONG C, XIA F, LI W, et al. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021. [S.l.] : Association for Computational Linguistics, 2021 : 2782 – 2794.

[6] van der MAATEN L, HINTON G. Visualizing Data using t-SNE[J/OL]. Journal of Machine Learning Research, 2008, 9(86) : 2579 – 2605. http://jmlr.org/papers/v9/vandermaaten08a.html.

[7] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space[C] // BENGIO Y, LECUN Y. 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings. 2013.

[8] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global Vectors for Word Representation[C] // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. [S.l.]: ACL, 2014: 1532–1543.

[9] DEVLIN J, CHANG M, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C] // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). [S.l.]: Association for Computational Linguistics, 2019: 4171–4186.

[10] CUI F, CUI Q, SONG Y. A Survey on Learning-Based Approaches for Modeling and Classification of Human-Machine Dialog Systems[J], 2021, 32(4): 1418–1432.

[11] ZHANG L, MA M, WANG P. Zero-shot Question Generation: Accelerate the Development of Domain-specific Dialogue System[C] //. [S.l.]: IEEE, : 123–128.

[12] YANG S, WANG Y, CHU X. A Survey of Deep Learning Techniques for Neural Machine Translation[J/OL]. CoRR, 2020, abs/2002.07526. https://arxiv.org/abs/2002.07526.

[13] GHANEM B, KAROUI J, BENAMARA F, et al. Idat at fire2019: Overview of the track on irony detection in arabic tweets[C] // Proceedings of the Forum for Information Retrieval Evaluation. 2019: 10–13.

[14] KHODAK M, SAUNSHI N, VODRAHALLI K. A Large Self-Annotated Corpus for Sarcasm[C] // Proceedings of the International Conference on Language Resources and Evaluation. 2018.

[15] WEBSTER K, RECASENS M, AXELROD V, et al. Mind the gap: A balanced corpus of gendered ambiguous pronouns[J]. Transactions of the Association for Computational Linguistics, 2018, 6 : 605 – 617.

[16] ARTSTEIN R, POESIO M. Inter-coder agreement for computational linguistics[J], 2008, 34(4) : 555 – 596.

[17] SIM J, WRIGHT C C. The kappa statistic in reliability studies: use, interpretation, and sample size requirements[J]. Physical therapy, 2005, 85(3) : 257 – 268.

[18] MISHRA A, DEY K, BHATTACHARYYA P. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network[C] // Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2017 : 377 – 387.

[19] KANT N, PURI R, YAKOVENKO N, et al. Practical text classification with large pre-trained language models[J]. arXiv preprint, 2018, arXiv:1812.01207.

[20] LIN S, HSIEH S. Sarcasm Detection in Chinese Using a Crowdsourced Corpus[C] // Proceedings of the Conference on Computational Linguistics and Speech Processing. 2016 : 299 – 310.

[21] ROSENTHAL S, FARRA N, NAKOV P. SemEval-2017 Task 4: Sentiment Analysis in Twitter[C] // Proceedings of the International Workshop on Semantic Evaluation. 2017 : 502 – 518.

[22] ÖHMAN E, PÀMIES M, KAJAVA K, et al. XED: A Multilingual Dataset for Sentiment Analysis and Emotion Detection[C/OL] // Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online) : International Committee on Computational Linguistics, 2020 : 6542 – 6552. https://aclanthology.org/2020.coling-main.575.

[23] SHANAHAN J G, ROMA N. Improving SVM Text Classification Performance through Threshold Adjustment[C] // LAVRAČ N, GAMBERGER D, BLOCKEEL H, et al. Machine Learning: ECML 2003. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003: 361 – 372.

[24] SABRI M, FIEGUTH P. A New Gabor Filter Based Kernel for Texture Classification with SVM[C] // CAMPILHO A, KAMEL M. Image Analysis and Recognition. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004: 314 – 322.

[25] KEERTHI S S. Generalized LARS as an Effective Feature Selection Tool for Text Classification with SVMs[C] // ICML '05: Proceedings of the 22nd International Conference on Machine Learning. New York, NY, USA: Association for Computing Machinery, 2005: 417 – 424.

[26] SU J, SHIRAB J S, MATWIN S. Large Scale Text Classification using Semisupervised Multinomial Naive Bayes[C] // GETOOR L, SCHEFFER T. Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011. [S.l.]: Omnipress, 2011: 97 – 104.

[27] DAI W, XUE G, YANG Q, et al. Transferring Naive Bayes Classifiers for Text Classification[C] // Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada. [S.l.]: AAAI Press, 2007: 540 – 545.

[28] WAHIBA B A, AHMED B E F. New Fuzzy Decision Tree Model for Text Classification[C] // GABER T, HASSANIEN A E, EL-BENDARY N, et al. Advances in Intelligent Systems and Computing, Vol 407: The 1st International Conference on Advanced Intelligent System and Informatics, AISI 2015, November 28-30, 2015, Beni Suef, Egypt. [S.l.]: Springer, 2015: 309 – 320.

[29] ASEERVATHAM S, GAUSSIER É, ANTONIADIS A, et al. Logistic Regression and Text Classification[G] // GAUSSIER É, YVON F. Textual Information Access: Statistical Models. [S.l.]: Wiley-ISTE, 2012: 61 – 84.

[30] TIMOTHY P J. maxent: An R Package for Low-memory Multinomial Logistic Regression with Support for Semi-automated Text Classification[J/OL]. R J.,

2012, 4(1) : 56.
https://doi.org/10.32614/rj-2012-007.

[31] PRANCKEVICIUS T, MARCINKEVICIUS V. Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification[J/OL]. Balt. J. Mod. Comput., 2017, 5(2).
https://doi.org/10.22364/bjmc.2017.5.2.05.

[32] ZHANG L, WANG S, LIU B. Deep learning for sentiment analysis: A survey[J]. Wiley Interdiscip. Rev. Data Min. Knowl. Discov., 2018, 8(4).

[33] LIU B. Sentiment analysis and opinion mining[J]. Synthesis lectures on human language technologies, 2012, 5(1) : 1 – 167.

[34] BATAA E, WU J. An Investigation of Transfer Learning-Based Sentiment Analysis in Japanese[C/OL] // KORHONEN A, TRAUM D R, MÀRQUEZ L. Proceedings of the 57th Conference of the Association for Computational Linguistics, (Volume 1: Long Papers), pages 4652–4657, Florence, Italy. [S.l.] : Association for Computational Linguistics, 2019.
https://doi.org/10.18653/v1/p19-1458.

[35] SCHMITT M, STEINHEBER S, SCHREIBER K, et al. Joint Aspect and Polarity Classification for Aspect-based Sentiment Analysis with End-to-End Neural Networks[C] // RILOFF E, CHIANG D, HOCKENMAIER J, et al. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium : Association for Computational Linguistics, 2018 : 1109 – 1114.

[36] TANG J, LU Z, SU J, et al. Progressive Self-Supervised Attention Learning for Aspect-Level Sentiment Analysis[C] // Proceedings of the 57th Conference of the Association for Computational Linguistics (Volume 1: Long Papers). Florence, Italy : Association for Computational Linguistics, 2019 : 557 – 566.

[37] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of Tricks for Efficient Text Classification[C] // Proceedings of the 15th Conference of the European Chapter

of the Association for Computational Linguistics: Volume 2, Short Papers. [S.l.]: Association for Computational Linguistics, 2017: 427–431.

[38] LUO F, LI P, YANG P, et al. Towards fine-grained text sentiment transfer[C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 2013–2022.

[39] BAO L, LAMBERT P, BADIA T. Attention and Lexicon Regularized LSTM for Aspect-based Sentiment Analysis[C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Florence, Italy: Association for Computational Linguistics, 2019: 253–259.

[40] LIANG B, DU J, XU R, et al. Context-aware Embedding for Targeted Aspect-based Sentiment Analysis[C] // KORHONEN A, TRAUM D R, MÀRQUEZ L. Proceedings of the 57th Conference of the Association for Computational Linguistics, (Volume 1: Long Papers). Florence, Italy: Association for Computational Linguistics, 2019: 4678–4683.

[41] JOSHI A, BHATTACHARYYA P, CARMAN M J. Automatic Sarcasm Detection: A Survey[J]. ACM Comput. Surv., 2017, 50(5): 73:1–73:22.

[42] GHOSH D, VAJPAYEE A, MURESAN S. A Report on the 2020 Sarcasm Detection Shared Task[C] // KLEBANOV B B, SHUTOVA E, LICHTENSTEIN P, et al. Proceedings of the Second Workshop on Figurative Language Processing, Fig-Lang, pages 1–11, Online. [S.l.]: Association for Computational Linguistics, 2020.

[43] MISHRA A, KANOJIA D, NAGAR S, et al. Harnessing Cognitive Features for Sarcasm Detection[C] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, (Volume 1: Long Papers), Berlin, Germany. [S.l.]: The Association for Computer Linguistics, 2016.

[44] ZHANG M, ZHANG Y, FU G. Tweet sarcasm detection using deep neural network[C] // Proceedings of COLING 2016, The 26th International Conference on Computational Linguistics: Technical Papers. 2016: 2449–2460.

[45] PENNINGTON J, SOCHER R, MANNING C D. GloVe: Global Vectors for Word Representation[C] // Empirical Methods in Natural Language Processing (EMNLP). 2014 : 1532 – 1543.

[46] TAY Y, LUU A T, HUI S C, et al. Reasoning with Sarcasm by Reading In-Between[C] // GUREVYCH I, MIYAO Y. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia : Association for Computational Linguistics, 2018 : 1010 – 1020.

[47] HAZARIKA D, PORIA S, GORANTLA S, et al. CASCADE: Contextual Sarcasm Detection in Online Discussion Forums[C] // BENDER E M, DERCZYNSKI L, ISABELLE P. Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA : Association for Computational Linguistics, 2018 : 1837 – 1848.

[48] REI M, BULAT L, KIELA D, et al. Grasping the Finer Point: A Supervised Similarity Network for Metaphor Detection[C] // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark : Association for Computational Linguistics, 2017 : 1537 – 1546.

[49] JANG H, JO Y, SHEN Q, et al. Metaphor Detection with Topic Transition, Emotion and Cognition in Context[C] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany : Association for Computational Linguistics, 2016 : 216 – 225.

[50] BIZZONI Y, LAPPIN S. Predicting Human Metaphor Paraphrase Judgments with Deep Neural Networks[C] // Proceedings of the Workshop on Figurative Language Processing. New Orleans, Louisiana : Association for Computational Linguistics, 2018 : 45 – 55.

[51] STOWE K, PALMER M. Leveraging Syntactic Constructions for Metaphor Identification[C] // Proceedings of the Workshop on Figurative Language Processing. New Orleans, Louisiana : Association for Computational Linguistics, 2018 : 17 – 26.

[52] MOSOLOVA A, BONDARENKO I, FOMIN V. Conditional Random Fields for Metaphor Detection[C] // Proceedings of the Workshop on Figurative Language Processing. New Orleans, Louisiana : Association for Computational Linguistics, 2018 : 121 – 123.

[53] MAO R, LIN C, GUERIN F. Word embedding and wordnet based metaphor identification and interpretation[C] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018 : 1222 – 1231.

[54] MAO R, LIN C, GUERIN F. End-to-end sequential metaphor identification inspired by linguistic theories[C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy : Association for Computational Linguistics, 2019 : 3888 – 3898.

[55] MAJUMDER N, PORIA S, PENG H, et al. Sentiment and sarcasm classification with multitask learning[J]. IEEE Intelligent Systems, 2019, 34(3) : 38 – 43.

[56] JIN N, WU J, MA X, et al. Multi-Task Learning Model Based on Multi-Scale CNN and LSTM for Sentiment Classification[J]. IEEE Access, 2020, 8 : 77060 – 77072.

[57] AKHTAR M S, GARG T, EKBAL A. Multi-task learning for aspect term extraction and aspect sentiment classification[J]. Neurocomputing, 2020, 398 : 247 – 256.

[58] YAN H, DENG B, LI X, et al. TENER: Adapting Transformer Encoder for Named Entity Recognition[J]. CoRR, 2019, abs/1911.04474.

[59] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All you Need[C] // GUYON I, von LUXBURG U, BENGIO S, et al. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. 2017 : 5998 – 6008.

[60] OUYANG X, WANG S, PANG C, et al. ERNIE-M: Enhanced Multilingual Representation by Aligning Cross-lingual Semantics with Monolingual Corpora[C]

// MOENS M, HUANG X, SPECIA L, et al. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. [S.l.] : Association for Computational Linguistics, 2021 : 27 – 38.

[61] MA R, PENG M, ZHANG Q, et al. Simplify the Usage of Lexicon in Chinese NER[C] // JURAFSKY D, CHAI J, SCHLUTER N, et al. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. [S.l.] : Association for Computational Linguistics, 2020 : 5951 – 5960.

[62] LI X, YAN H, QIU X, et al. FLAT: Chinese NER Using Flat-Lattice Transformer[C] // JURAFSKY D, CHAI J, SCHLUTER N, et al. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. [S.l.] : Association for Computational Linguistics, 2020 : 6836 – 6842.

[63] ZHENG C, CAI Y, XU J, et al. A Boundary-aware Neural Model for Nested Named Entity Recognition[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). 2019 : 357 – 366.

[64] TAN C, QIU W, CHEN M, et al. Boundary Enhanced Neural Span Classification for Nested Named Entity Recognition[C] // The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. [S.l.] : AAAI Press, 2020 : 9016 – 9023.

[65] LIN H, LU Y, HAN X, et al. Sequence-to-Nuggets: Nested Entity Mention Detection via Anchor-Region Networks[C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. [S.l.] : Association for Computational Linguistics, 2019.

[66] JU M, MIWA M, ANANIADOU S. A Neural Layered Model for Nested Named Entity Recognition[C] // 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 : 1446 – 1459.

[67] FU Y, TAN C, CHEN M, et al. Nested Named Entity Recognition with Partially-Observed TreeCRFs[C] // Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. [S.l.] : AAAI Press, 2021 : 12839 – 12847.

[68] YAN H, GUI T, DAI J, et al. A Unified Generative Framework for Various NER Subtasks[C] // ZONG C, XIA F, LI W, et al. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021. [S.l.] : Association for Computational Linguistics, 2021 : 5808 – 5822.

[69] GHANEM B, KAROUI J, BENAMARA F, et al. IDAT at FIRE2019: Overview of the Track on Irony Detection in Arabic Tweets[C] // MAJUMDER P, MITRA M, GANGOPADHYAY S, et al. FIRE '19: Forum for Information Retrieval Evaluation, Kolkata, India, December, 2019. [S.l.] : ACM, 2019 : 10 – 13.

[70] KHODAK M, SAUNSHI N, VODRAHALLI K. A Large Self-Annotated Corpus for Sarcasm[C] // CHAIR) N C C, CHOUKRI K, CIERI C, et al. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan : European Language Resources Association (ELRA), 2018.

[71] WEBSTER K, RECASENS M, AXELROD V, et al. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns[J]. Trans. Assoc. Comput. Linguistics, 2018, 6 : 605 – 617.

[72] ARTSTEIN R, POESIO M. Inter-Coder Agreement for Computational Linguistics[J]. Comput. Linguistics, 2008, 34(4) : 555 – 596.

[73] ESCOBAR-PÉREZ J, CUERVO-MARTÍNEZ Á. Validez de contenido y juicio de expertos: una aproximación a su utilización[J]. Avances en medición, 2008, 6(1): 27–36.

[74] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory[J/OL]. Neural Comput., 1997, 9(8): 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

[75] TURC I, CHANG M-W, LEE K, et al. Well-Read Students Learn Better: On the Importance of Pre-training Compact Models[J]. arXiv preprint arXiv:1908.08962v2, 2019.

[76] BENGIO Y, SIMARD P Y, FRASCONI P. Learning long-term dependencies with gradient descent is difficult[J/OL]. IEEE Trans. Neural Networks, 1994, 5(2): 157–166. https://doi.org/10.1109/72.279181.

[77] ELKAN C. The foundations of cost-sensitive learning[C] // International joint conference on artificial intelligence: Vol 17. 2001: 973–978.

[78] BATISTA F, RIBEIRO R. Sentiment Analysis and Topic Classification based on Binary Maximum Entropy Classifiers[J]. Proces. del Leng. Natural, 2013, 50: 77–84.

[79] ALTMAN N S. An introduction to kernel and nearest-neighbor nonparametric regression[J]. The American Statistician, 1992, 46(3): 175–185.

[80] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning internal representations by error propagation[R]. [S.l.]: California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[81] REZATOFIGHI H, TSOI N, GWAK J, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C] // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 658–666.

[82] YU J, BOHNET B, POESIO M. Named Entity Recognition as Dependency Parsing[C/OL] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online : Association for Computational Linguistics, 2020 : 6470 – 6476.
https://aclanthology.org/2020.acl-main.577.

[83] WANG W, PAN S J, DAHLMEIER D. Memory networks for fine-grained opinion mining[J]. Artificial Intelligence, 2018, 265 : 1 – 17.

[84] CHO K, van MERRIENBOER B, BAHDANAU D, et al. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches[J/OL]. CoRR, 2014, abs/1409.1259.
http://arxiv.org/abs/1409.1259.

[85] MACQUEEN J, OTHERS. Some methods for classification and analysis of multivariate observations[C] // Proceedings of the fifth Berkeley symposium on mathematical statistics and probability : Vol 1. 1967 : 281 – 297.

致　　谢

三年的研究生时光转瞬即逝，回首往事，我在南京大学度过的时光历历在目。研究生的三年，也是疫情肆虐的三年，这让我对研究生时光更增添了一份不一样的记忆。临别之际，对于陪伴我三年的校园，以及在校园中遇到的人与事，我的心中充满感慨，这将成为我人生历程中难以忘却的宝贵财富。此刻，我由衷地感谢在研究生生涯中给予我关心、支持和帮助的师长与同伴。

首先，我要感谢我的导师申富饶老师。申老师治学严谨，对科研有一直坚守的品味，教导我从问题出发进行独立思考，做有价值的研究工作。同时，申老师在科研方向给予了我极大的自由度和强有力的支持。犹记得，当时在我即将放弃目前的研究方向时，是申老师鼓励我，让我坚持下去，并且在后面的过程中，给予我足够的支持。申老师的科研理念与对待科研的严谨态度对我产生了极大的影响，让我不再为了论文而研究，而是因为问题而研究。申老师对于科研的追求和纯粹，让我从内心深处受到启发，并也以此要求自己。此外，申老师坚持每周与组内每一位同学进行面对面的讨论交流，并组织讨论班分享彼此的研究内容。申老师的辛苦付出，让我在科研过程中，不断突破自我，取得进步。

其次，我要感谢赵健老师。赵老师多次为我们分享论文写作的经验，并逐字逐句地帮我们检查修改待投稿的英文论文，对于我们英文论文写作给予了很大帮助。同时，赵老师与我们共同参与讨论班，为我们的研究工作提出许多有价值的意见与建议。此外，南京大学计算机系的老师为我们开设了丰富的专业课程，为我们进行深入的科研打下坚实的基础。

接着，我要感谢陪伴我度过研究生生涯的同学。感谢 RINC 研究组的同学，无论是在讨论班，还是私下的交流中，都热心为我的科研与学习提供了许多帮助。同时，友好融洽的实验室关系为我的生活带来了许多欢乐与温暖。感谢张雅楠和赖碧兰同学对我的毕业论文提出意见。感谢刘晓涛同学在我研究过程中给予的帮助。感谢我的室友，在生活中给予我非常多关心与帮助，以及室友们在一起打比赛过程中对我的帮助。

最后，我要感谢我的家人和女友。他们为我提供了最坚强的后盾，良好的

生活保障及坚实的精神支柱使我能够将精力更多地投入到科研与学习之中，在面对困难时不畏惧，在面对压力时不放弃，顺利完成学业。

简历与科研成果

基本信息

严骅，男，汉族，1996 年 12 月出生，福建省龙岩人。

教育背景

2019 年 9 月 — 2022 年 6 月 南京大学人工智能学院 硕士

2015 年 9 月 — 2019 年 6 月 南开大学软件学院 本科

攻读硕士学位期间完成的学术成果

1. **Hua Yan**, Feng Han, Junyi An, Weikang Xiao, Jian Zhao, Furao Shen, "SASICM: A Multi-Task Benchmark For Subtext Recognition," *arXiv* preprint arXiv:2106.06944(2021).
2. **Hua Yan**, Suhan Guo, Junyi An, Feng Han, Jian Zhao, Furao Shen, "Subtext Recognition: Teach Machine To Read Between The Lines" Coling2022 under review.

攻读硕士学位期间完成的专利成果

1. 李俊, **严骅**，刘晓涛. "一种基于自然语言处理技术的网页文本内容的分类方法"(202110718603.5)

攻读硕士学位期间参与的科研课题

1. 国家自然科学基金项目 "基于深度感知增量式联想记忆神经网络的信息融合系统研究"（项目编号：61876067，课题年限 2019 年 1 月 — 2022 年 12 月），负责自然语言处理相关问题的研究。

学位论文出版授权书

本人完全同意《中国优秀博硕士学位论文全文数据库出版章程》（以下简称“章程”），愿意将本人的学位论文提交“中国学术期刊（光盘版）电子杂志社”在《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》中全文发表。《中国博士学位论文全文数据库》、《中国优秀硕士学位论文全文数据库》可以以电子、网络及其他数字媒体形式公开出版，并同意编入《中国知识资源总库》，在《中国博硕士学位论文评价数据库》中使用和在互联网上传播，同意按“章程”规定享受相关权益。

作者签名：_____

2022 年 05 月 20 日

论文题名	句子级与片段级潜台词分析研究				
研究生学号	MG1937028	所在院系	人工智能学院	学位年度	2022
论文级别	☑ 硕士 ☐ 硕士专业学位 ☐ 博士 ☐ 博士专业学位 （请在方框内画勾）				
作者 Email	williamyh@smail.nju.edu.cn				
导师姓名	申富饶 教授				

论文涉密情况：

☑ 不保密

☐ 保密，保密期：(_____年_____月_____日 至 _____年_____月_____日)

注：请将该授权书填写后装订在学位论文最后一页（南大封面）。