硕士学位论文

中文成语表征学习及其应用

Research on Chinese Idiom Representation Learning and Its Application

北京交通大学

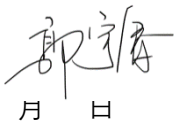
2021 年 5 月

学位论文版权使用授权书

本学位论文作者完全了解北京交通大学有关保留、使用学位论文的规定。特授权北京交通大学可以将学位论文的全部或部分内容编入有关数据库进行检索，提供阅览服务，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。学校可以为存在馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：　　导师签名：



签字日期：　　年　月　日　　签字日期：　　年　月　日

学校代码：10004 密级：公开

北京交通大学

硕士学位论文

中文成语表征学习及其应用

Research on Chinese Idiom Representation Learning and Its Application

作者姓名：李想 学　　号：18120100

导师姓名：郭宇春 职　　称：教授

学位类别：工　学 学位级别：硕士

学科专业：通信与信息系统 研究方向：自然语言处理

北京交通大学

2021 年 5 月

致谢

本论文的相关研究工作是在我的导师郭宇春老师，以及陈一帅老师的耐心指导下完成的。郭老师和陈老师严谨治学的态度，平易近人的性格的陪伴我度过了研究生生涯，深深地鼓舞着我不断克服学习和生活中的困难。本文从选题到研究方向，两位老师都给予了细心的指导，为我修改各种报告和答辩 PPT，提出许多针对性意见。除了在学业上的帮助，两位老师还在生活中无时无刻地影响着我，其对生活和研究的热爱，以及思想中迸发的火花，都启迪了我的不断成长和进步，在此我向两位老师致以最诚挚的感谢。

在三年实验室的学习经历中，孙欢、戚余航、王珍珠、曹中等同学对我论文中代码实现部分给予了无私的帮助，同我一起探讨研究中相关的问题，在此向他们表达真挚的谢意。

最后，感谢我的家人们在我成长过程中给予我的支持，他们二十多年来默默付出，我才能顺利完成学业，走向社会，成为建设社会的有用之才。

摘要

随着深度学习的兴起，自然语言处理在中文领域快速发展，其中文本表征是不可或缺的基础编码层。成语在书面和口语中使用频繁，在中文表意中有着非常重要的作用，地位不可替代。因此，高效的成语表征对中文自然语言处理的进一步发展至关重要。

成语是中文独特的语言现象，它固定的四字结构，形式简洁，内容丰富，带来了两大特性：非语义合成性和意义整体性，即：它的意义不能简单通过字的含义相加，而是一个整体。这两个特点导致目前主流的词级别和字级别中文文本表征方法不适合直接应用于中文成语。

为了有效对成语进行表征，本文提出基于释义增强的中文成语多粒度表征模型，并基于完形填空式的中文阅读理解任务验证表征效果，最后将其应用于高考语文成语试题中，获得了较好的效果。

本文贡献如下：

（1）本文提出了两种表征模型。1）字词融合的上下文表征模型。为了实现字词表征的完美融合，本文设计了两种字词向量对齐方法，解决字词向量对齐问题；提出了三种融合方式，对字词之间的交互方式进行建模。2）基于释义增强的成语表征模型。为了完成对释义中不同成分的有效筛选，本文设计了独特的注意力机制，解决了词向量无法对成语进行有效表征、成语字信息会对词信息造成混淆的两个问题。在真实中文机器阅读理解任务上的实验表明：本文模型能够改进目前主流的 BiLSTM、AR 和 SAR 阅读理解模型的性能，最高能够提升 9.5%，证明了该方法的有效性和通用性。

（2）本文通过具体案例的量化分析，发现上述模型获得的相似成语的表征之间的欧式距离更大，余弦相似度更小，证明了本文提出的成语表征模型具有比基线模型更强的相似成语辨别能力，是一个通用的表征模型，具有广泛的应用价值。

（3）通过收集数据，本文建立了一个高考语文试卷中与成语相关的试题数据集，将上述模型应用于高考语文成语试题的解题任务中。实验结果表明：本文提出的模型能够很好地解决高考成语试题的解题工作，在测试集中准确率达到 75.9%，远高于考生平均水平 66.7%。

图 10 幅，表 19 个，参考文献 53 篇。

关键词： 自然语言处理；成语表征；中文机器阅读理解

ABSTRACT

With the rise of deep learning, natural language processing is developing rapidly in the Chinese field, where text representation is an indispensable basic encoding layer. Idioms are frequently used in written and spoken language. They play a very important role in Chinese ideology and their status is irreplaceable. Therefore, efficient idiom representation is crucial to the further development of Chinese natural language processing.

Idioms are a unique language phenomenon in Chinese. Its fixed four-character structure, simple form and rich content bring two major characteristics: non-compositionality and meaning integrity, that is, its meaning cannot be simply added by the meaning of the characters, but as a whole. These two characteristics make the current mainstream word-level and character-level representation methods not suitable for direct application to Chinese idioms.

In order to effectively represent idioms, we propose a multi-granularity representation model for Chinese idioms based on the definition-augmented embedding, and a cloze-style Chinese reading comprehension task to verify the representation effect. Finally, it is applied to the college entrance examination Chinese idiom test questions, and achieves good results.

The contributions of this paper are as follows:

1) We propose two representation models. 1) Contextual representation model based on the mixed embedding of characters and words. In order to achieve the perfect integration of characters and words, we design two word vector alignment methods to solve the alignment problem and three fusion methods to model the interaction between characters and words. 2) Idiom representation model based on the definition-augmented embedding. In order to complete the effective screening of the different components in the definition, a unique attention mechanism is designed to solve the two problems that the word vector cannot effectively represent the idiom and the character information will cause the confusion of the word information. Experiments on real Chinese machine reading comprehension tasks show that the model in this paper can improve the performance of the current mainstream BiLSTM, AR and SAR reading comprehension models by up to 9.5%, which proves the effectiveness and versatility of our method.

2) Through the quantitative analysis of specific cases, we find that the Euclidean distance between the representations of similar idioms obtained by the above model is

larger and the cosine similarity is smaller, which proves that the idiom representation model proposed in this paper has stronger ability to distinguish similar idioms than the baseline model, which means our model is s a general characterization model with a wide range of application value.

3) We collect data and establishes a data set of test questions related to idioms in the college entrance examination Chinese test paper, and applies the above model to solve the questions. The experimental results show that the model proposed in this paper can solve the college entrance examination idiom questions very well. The accuracy rate in the test set is 75.9%, which is much higher than the average level of 66.7% of the human.

10 figures, 19 tables, and 53 reference articles are contained in the dissertation.

KEYWORDS: Natural language processing; Idiom representation; Chinese machine reading comprehension

目录

1 引言

1.1 研究背景和意义

成语是中文特有的一种语言文化现象，它有着固定的四字结构，内容简练、含义丰富，是一种语言文化的浓缩产物，深刻体现了中国文化的重意不重形的特点。大部分成语来源自典故，成语的运用需要有深厚的文化常识积累。深入理解成语对机器阅读理解、作文批改、网络舆情分析等中文自然语言处理任务都有着重要的作用。

对成语进行正确的数字化表征，获得成语的嵌入式表征，是中文成语机器阅读理解任务中的核心任务之一。文本的表征是自然语言处理基本任务，其目的是将非结构化的字符形式的文本编码为结构化的数字形式的向量，便于计算机处理。如何将文本中蕴含的语义信息、语法信息和结构信息挖掘出来是文本表征要解决的基本问题，得到的向量表征除了可以表征基础的语义单元，还要能够对更长的文本进行语言建模，即从字、词、句子、段落、再到文档。文本表征将文本切分为有意义的语言单位的序列，大部分英文单词都是有具体意义的语言单位，英文的书写方式使得空格就是其自然的分界符[1]，而现代汉语的绝大部分语义是由词构成，且词之间没有明显的间隔，分词的准确性影响着下游任务的表现，如机器阅读理解、文本分类、命名实体识别、文本摘要等。成语中大部分字使用了古汉语的意义，一字即一词，这就需要在处理成语时引入其字的信息，而引入字的信息又会对一些容易望文生义的成语产生错误的理解。因此如何对成语进行正确的数字化表征，获得成语的嵌入式表征，是中文成语机器阅读理解任务中的重要一环。

准确的中文成语表征对研究和开发各种中文成语相关的自然语言应用至关重要。以人工智能语文教育为例，随着人工智能的发展，在线教育平台可以为学生提供更加丰富多样的服务，如智能导学、习题推荐、作文批改等，教育的数字信息化有利于解决教育资源紧缺的问题，提高教育的质量、效率，最小化学生的学习时间[2]。例如作文练习，不同于其他有固定答案的习题，其具有语法、语义、写作技巧等多重评判标准，而中文灵活多样的语言形式，重意不重形的写作特点，以及大量成语典故的应用[3]，使得中文作文的机器批改成为难题。在智能语文教学中，成语的深度表征是一个关键性的基础工作。除了智能语文教学，成语的准确表征还对网络舆情分析[4]、写作校对、智能辅助工具等方面也有重要的应用。

目前，自然语言处理领域中关于中文成语表征的研究还是一个新的课题，因此，

本课题拟利用自然语言处理技术和深度学习工具，以机器阅读理解为目标任务，探究不同粒度的成语表征方式及其组合，建立中文成语的深度表征。这一研究具有很强的实用价值和经济意义，既能增强中文阅读理解任务中成语表征的质量，又可以应用于其他的中文自然语言处理任务，如舆情分析、文本生成、情感分析、语义消歧等。同时本课题提出的表征方法也适用于其他中文语言现象，如歇后语、对联、诗词等。

1.2 国内外研究现状

目前有大量关于中文自然语言处理的研究工作。下面将首先对文本表征和阅读理解相关研究进行介绍，然后介绍中文成语的研究现状。

1.2.1 文本表征模型相关研究

在中文中，最小语义单位一般是由词构成的，所以中文的表征大都集中于词级别，但是词级别的表征信息是粗粒度的，混有许多分词错误引入的噪声。词向量模型的目标是得到低维、稠密的向量表示，2013 年谷歌公司的 Mikolov 等人提出的 Word2Vec[5]，其基于基本的假设“一个词的含义可以通过其上下文获得”，后来，Mikolov 等人又延续 Word2Vec 的思想提出 Doc2Vec[6]来表征整个文档。另外一个著名的词向量模型 Glove（Global Vectors for Word Representation）[7]和 Word2Vec 一样，本质上是基于词的共现矩阵来进行。但是这种思想忽略了词的顺序，而且无法表征未知词，尤其是对于中文这种合成词较多的语言，Wang 等人[8]基于大规模语料库系统地探讨了中文的合成词特性。

有了词向量后可以得到句子的表征，简单的合成方法如 FastText[9]针对于文本分类任务，对组成句子的所有词向量相加后求取平均值就可以作为句子的表征，使用了浅层的网络极大节约了计算资源。此外，针对于句子和文档的建模工作，最常见的网络结构是卷积神经网络（Convolutional Neural Networks, CNN）和循环神经网络（Recurrent Neural Network, RNN），如 Kim 等人使用卷积神经网络来进行文本的表征[10]，但卷积神经网络只考虑有限固定窗口内的词向量，对于长距离的依赖信息和词序无法有效建模，因此一些隐藏语义信息无法捕捉。CNN 擅长提取局部特征，而 LSTM 擅长捕捉长距离信息，Lai 等人尝试结合两种网络的优点，提出 RCNN 用于文本建模[11]。

另一类工作针对中文词语的特性，考虑进了字的信息对词向量进行信息增强。Li S 等人[12]中使用大量的语料，针对词级别和字级别分别预训练了不同粒度的中

文字和词的表征。一些研究者[13]-[20]在预训练模型在迁移到中文时，均使用了字级别的输入信息，部分预训练模型在预训练阶段引入了词的信息，使其得到的字级别表征隐含了词的信息。中文的字信息对于粗粒度的词信息有着重要的信息补充修正作用，一些研究者将字的表征信息融合到词表征向量中进行中文文本分类[21][22]，还有一些研究者引入了介于字和词之间的“子词（subword）”的概念[23][24]，利用子词构建中文中最小语义单元。

总的来说，中文自然语言处理的深度表征方法借鉴了英文深度表征的主要思想，同时针对中文字组成词的特点，进行了有针对性的模型设计和调优。但是这些工作主要针对现代汉语中普通的词汇，但中文成语的语义特性和普通的中文词汇相差较大，因此不适合直接应用于成语的表征。而问题上下文则采用了白话文的现代汉语，与此类工作契合，本文拟沿此思路展开上下文相关的研究工作，具有很强的可行性。

1.2.2 阅读理解模型相关研究

注意力机制（Attention Mechanism）被认为是一种有效选择信息的方式，可以过滤掉大量与目标无关的信息，在自然语言处理领域最先在机器翻译任务中提出[25]，解决 seq2seq（sequence-to-sequence）模型在编码过程中把源序列映射成固定大小的向量时存在信息损失的情况。随后注意力机制被推广到各种自然语言处理的任务中，文本表征任务也不例外。主要有两种主流的注意力机制：分层注意力模型（Hierarchical Attention）[26]和自注意力模型（Self-Attention）[27]。其中分层注意力模型是输入对输出的权重，而自注意力模型是自己的对自己的权重，即文章中的不同词之间的权重，这样做是为了充分建立句子之中不同词语之间语义及语法联系。自注意力机制的提出对模型文本表征和语言建模能力有了重大的提升，之后许多高性能预训练模型都是基于自注意力机制，如 Google BERT[13]、OpenAI GPT[14]、OpenAI GPT2[15]、Transformer-XL[16]、Baidu ERNIE[17]、Tsinghua ERNIE[18]、MASS[19]、XLNet[20]，这些预训练模型突破了许多自然语言处理任务的性能瓶颈。这些预训练模型大都基于英文，一些学者将其应用到了中文领域，如 Cui Y 等人基于全词掩码构建了就中文领域的 BERT 预训练模型[28]。本文的相关模型也基于注意力机制开展研究。

还要许多工作针对于中文机器阅读理解任务，构建了中文阅读理解相关的数据集。之前许多公开数据集都是基于英文的，而英文语法与中文相差较大，因此一些研究者致力于构造中文相关机器阅读理解数据集。如 He W 等人构建了 DuReader 中文阅读理解数据集[29]，该数据集基于百度搜索和百度知道的真实数据，而且答

案均为人工手动生成，包含了丰富问题类型；Sun K 等人构建了第一个多项选择题形式的中文机器阅读理解数据集[30]，且该数据集的问题设置较难，许多问题的解答需要相关的先验知识；Cui Y 等人构建了 CMRC-2017 数据集[31]，该数据集面向儿童读物领域，采用了完形填空的形式，文章中包含较多拟人化的动植物实体词，更具有挑战性，同时他们还构建了 CFT 数据集[32]，也是基于完形填空式的问题类型，但数据领域涵盖了新闻（《人民日报》）和儿童读物。成语的学习中一大难点就是相似成语的辨析，因此本文使用完形填空类型的中文机器阅读理解任务来验证成语表征的有效性，即在一个位置设置多个与正确答案相似的成语，以更好地对相似成语进行辨析。

1.2.3 中文成语的研究现状

目前针对中文成语的研究主要集中于成语的数据库构建以及基本任务，还没有针对成语表征的具体研究。一些研究者致力于构建成语相关数据集，一些研究者构建了中文成语的知识图谱[33][34]；还有一些研究者构建了中文成语机器阅读理解数据集[35][36]，通过完形填空的形式，使机器能够辨别相似成语之间的细微差距；Liu Y 等人将成语视为一个推荐问题[37]，构建了成语推荐数据集，旨在根据文本的语义信息，为每段文本推荐一个适合的成语；Grace 等人基于维基百科在线资源[38]，提出自动识别成语、习语等惯用语词条的模型，并提供了高质量的成语标签数据集。

在中文机器阅读理解方面，一些研究者根据成语的特性提出了相关的阅读理解模型。Long 等人基于同义成语构建了同义词图[39]，结合图注意力神经网络和门机制对其进行编码，并用于解决中文成语阅读理解；Tan 等人基于预训练模型 BERT 提出成语阅读理解任务双重嵌入式表征模型[40]；Liu 等人提出一种基于成语感知的分布式语义模型[41]，在理解句子包含的成语的基础上构建句子的表征，并构建了一个成语感情分类数据集；Shao Y 等人关注于中文成语的机器翻译任务[42]，引入了针对于中文成语的评估方法，并构建了中文成语机器翻译黑名单，可以有效地识别出机器翻译中的错误。

上述工作均针对于成语的特性研究相关的下游任务，还没有针对成语上游的表征任务的具体研究。本文拟从基础的编码层开展研究，探究如何对成语进行有效表征，这一方向具有重要的研究价值和实际意义。

1.3 研究内容及主要贡献

四字成语是汉语成语的主要形式。所以本文对成语的分析，主要集中在四字成

语。相较于普通文本文的表征，成语一个重要特性是成语的非语义合成性[43]以及其意义的整体性。非语义合成性是指一些成语如果拆开来分析得到的解释可能和其实际意义不同甚至完全相悖，如成语“不刊之论”，如果拆开使用现代汉语解释很容易误以为“不能刊登的言论”，然而其实际意义是“比喻不能改动或不可磨灭的言论，用来形容文章或言辞的精准得当”；意义整体性是指在其构成成分的意义基础上进一步概括出来的整体意义，如“亡羊补牢”，其意义为“比喻出了问题以后想办法补救，免得以后继续受损失”，是在故事中提取抽象出来的整体意义。所以如何在不失其整体意义的基础上引入多粒度的信息，从不同的粒度构建成语的表征，是个亟待解决的问题。

从上述研究现状中可以看出，现有针对成语的表征主要有两种方案：基于词信息表征和基于字词融合的方法。两种方案都存在明显的问题。基于词向量的方法对于频率较低的词无法有效表征，且无法解决一词多义的问题；而字词融合的方法则受限于成语的历史性，即成语中的字大部分沿用了古汉语的释义，而训练字信息的语料库一般都基于现代汉语[44]，其语义特性和古汉语有较大出入，使用组成成语的字向量对其词向量进行信息增强反而会使其性能降低。

因此，本文将成语与上下文进行分开编码，并引入了成语的释义对其进行信息增强。首先，对于使用现代汉语释义的上下文和成语释义，采用字词融合的方法对其进行表征；然后，对于成语则采用词向量进行表征，并使用其释义对其词向量进行信息补充；最后，在机器阅读理解的下游任务中验证表征的性能。

本文工作难点主要有：

（1）成语相对于普通文本其语法功能多样，可能出现在句子中的任何位置，其意义具有较强的抽象性，这会对模型效果产生干扰；

（2）由于成语中大部分的字沿用了古汉语的释义，而目前字信息的获取主要通过现代汉语训练得到，字级别的表征可能存在现代汉语和古代汉语杂糅的现象，会对得到的成语表征产生干扰；

（3）通过收集普通高等学校招生全国统一考试（简称为高考，下同）语文试卷中与成语相关的试题数据，验证模型在实际应用中的表现；

（4）现有主流模型大都针对英文，而中文结构灵活多变，不像英文有着明显的语法结构信息，模型在中文的情形下可能表现不佳。

本文的主要贡献为：

（1）本文通过分开编码实现了本文各个部分的有效表征。首先，针对组成成语的字的沿用古汉语释义的特征，引入了其现代汉语释义，使用字词融合的方式对上下文文本和成语释义进行表征，然后使用得到的释义表征对成语词向量进行信息补充；其次，对于采用现代汉语释义的上下文文本，则采用字词融合的方式，使

用字信息对词信息进行有效的信息补充，并探究了多种字词信息融合的方法。

（2）基于完形填空式的中文阅读理解任务以验证表征的有效性，通过多组消融实验，探究最佳的表征方式，实验结果表明，本文提出的“字词融合+释义增强”表征模型的性能表现相较于基线模型性能最高提升了 9.5%。

（3）针对成语的实际应用场景，本文收集了高考语文试卷中成语相关的试题作为额外的测试集以验证模型的实际应用效果，实验结果表明，本文提出的模型在高考测试集中准确率达到了 75.9%，远远高于考生的平均水准 66.7%。

（4）本文提出的中文成语表征方法可以扩展到其他类似的中文语言中，进行有效的表征，如对联、古诗词等。

1.4 论文组织结构

本文整体的组织结构如下：

第二章为本文研究工作中相关的知识背景和技术介绍。包括自然语言处理在中文的发展，词向量的相关知识，以及本文所使用的相关开发平台和深度学习框架。

第三章对本文所使用的成语数据集进行介绍，包括数据集的基础统计分析和任务定义，并实验验证基线模型。

第四章提出了基于字词融合的神经网络模型。该模型针对于现代汉语，对文章的上下文进行表征，并通过具体实验进行验证。

第五章提出了基于成语词信息和释义融合的神经网络模型。该模型针对于成语建模，使用成语释义对成语词向量进行信息增强，并通过消融实验进行结果分析，最后进行结果的可视化分析。

第六章通过具体的案例对模型结果进行分析，通过具体度量标准对效果提升进行量化分析；然后，收集高考成语相关的试题数据作为额外的测试集以验证模型的实际应用效果。

第七章对全文内容进行了总结并对未来工作方向进行相关介绍。

2 技术背景

本章介绍研究工作相关的技术背景，包括中文自然语言处理，中文文本表征和机器阅读理解相关算法，同时介绍本文实验中用到的开发平台。

2.1 自然语言处理

自然语言处理是(Natural Language Processing，NLP)一门融语言学、计算机科学、数学、认知学、逻辑学于一体的典型边缘交叉学科，是人工智能（Artificial Intelligence, AI）的一个子领域。自然语言是人类智慧的结晶，自然语言处理是人工智能中最为困难的问题之一，它是能够让人类与智能机器进行沟通交流的重要技术手段。

自然语言处理是指利用人类交流所使用的自然语言与机器进行交互通讯的技术。通过人为的对自然语言的处理，使得计算机对其能够可读并理解。该领域涉及到的自然语言指的是人们日常使用的语言，所以它与语言学的研究有着密切的联系，但又有重要的区别：自然语言处理并不是一般地研究自然语言，而在于研制能有效地实现自然语言通信的计算机系统，特别是其中的软件系统，因而它是计算机科学的一部分，即把计算机作为语言研究的强大工具，在计算机的支持下对语言信息进行定量化的研究，并提供可供人与计算机之间能共同使用的语言描写。

自然语言处理包括自然语言理解(Natural Language Understanding, NLU)和自然语言生成(Natural Language Generation, NLG)两部分：

（1）自然语言理解目的是使计算机理解自然语言，重在理解。具体来说，就是理解语言、文本等，提取出有用的信息，用于下游的任务。典型的自然语言理解任务包括：自然语言结构化，比如分词、词性标注、句法分析等；表征学习，将字、词、句子进行数学化表示，获得其嵌入式表征(Embedding)；信息提取，如信息检索、文本匹配、命名实体提取、关系抽取、事件抽取等。

（2）自然语言生成是研究使计算机具有人一样的表达和写作的功能。即能够根据一些关键信息及其在机器内部的表达形式，经过一个规划过程，来自动生成一段高质量的自然语言文本。典型的自然语言生成任务包括：机器翻译、人机对话、知识问答、文本摘要等。

自然语言处理的发展可大致分为两个阶段：统计自然语言处理和神经网络自然语言处理。

（1）统计自然语言处理：基于统计学的机器学习(Machine learning)。主要思路

是利用带标注的数据，基于人工定义的特征建立机器学习系统，并利用数据经过学习确定机器学习系统的参数。运行时利用这些学习得到的参数，对输入数据进行解码，得到输出。传统机器学习算法例如支持向量机（Support Vector Machine, SVM）、线性回归（Linear Regression, LR）等，对映射到高维空间的文本特征进行处理，大部分应用在文本分类、情感分析、机器翻译、搜索引擎等。

（2）神经网络自然语言处理：基于神经网络（Neural Network）。神经网络是一种运算模型，由大量的节点（或称神经元）之间相互联接构成。每个节点代表一种特定的输出函数，称为激励函数（Activation Function）。每两个节点间的连接都代表一个对于通过该连接信号的加权值，称之为权重，这相当于人工神经网络的记忆。网络的输出则依网络的连接方式，权重值和激励函数的不同而不同。而网络自身通常都是对自然界某种算法或者函数的逼近，也可能是对一种逻辑策略的表达。典型的神经网络包括前馈神经网络（Feedforward Neural Network, FNN）、卷积神经网络（Convolutional Neural Networks, CNN）、长短期记忆网络（Long Short-Term Memory, LSTM）。而 Transformer 模型的提出标志着自然语言处理进入了新的发展阶段，Transformer 摆脱了自然与处理任务对于 LSTM 的依赖，使用了自注意力（self-attention）的方式对上下文进行建模，实现了并行运算，提高了训练和推理的速度，此后各种基于此结构的预训练模型极大提高了自然语言处理各类任务的性能。

2.2 中文文本表征

2.2.1 独热词表征

独热编码即 One-Hot 编码，又称一位有效编码。给定词表集合 $V = \{w_1, w_2, \ldots, w_{|V|}\}$，词表大小为|V|，独热编码将词$w_i$映射为一个维度为|V|的向量，其中仅第 i 维为 1，其他维度为 0。

从本质上讲，独热词表征法将每个单词映射为其在词汇表的索引，这对于存储和计算非常有效，但是信息存储密度非常低。而且这种方法却忽略了词义和句法信息，无法捕捉词之间的关联性。此外，独热词表征将每个词嵌入|V|维向量中，该向量仅适用于固定的词汇表，而在现实世界中新词不断出现，因此该方法也缺乏灵活性。

2.2.2 Word2Vec

分布式词表征的目的在于将词嵌入到一个连续密集（dense）实值向量，称为嵌

入式词向量（Word Embedding）。这里的密集指的是一个概念用多个维度表示，同时一个维度涉及多个概念的表示。分布式词表征中最著名的模型是 Word2Vec，即 Word-to-vector，Word2Vec 是用一个一层的神经网络把 one-hot 形式的稀疏词向量映射称为一个低维的密集向量的过程，包括两个重要的模型：CBOW 模型(Continuous Bag-of-Words Model)与 Skip-Gram 模型。

（1）CBOW

CBOW 就是根据某个词前面的 c 个连续词和后面的 c 个连续词来计算某个词出现的概率，计算公式为：

$$P\left(w_i \middle| w_{j(|j-i|\leq c, j\neq i)}\right) = Softmax(\boldsymbol{M}(\sum_{|j-i|\leq c, j\neq i} w_j)) \tag{2-1}$$

其中$P\left(w_i \middle| w_{j(|j-i|\leq c, j\neq i)}\right)$为给定前后 c 个连续词，中心词为$w_i$的概率，c 为上下文窗口大小。$\boldsymbol{M}$为权重矩阵，维度为$R^{|V|\times m}$，|V|为词表大小，m 为词向量的维度，矩阵每一行代表一个词向量。模型结构如图 2-1 所示：

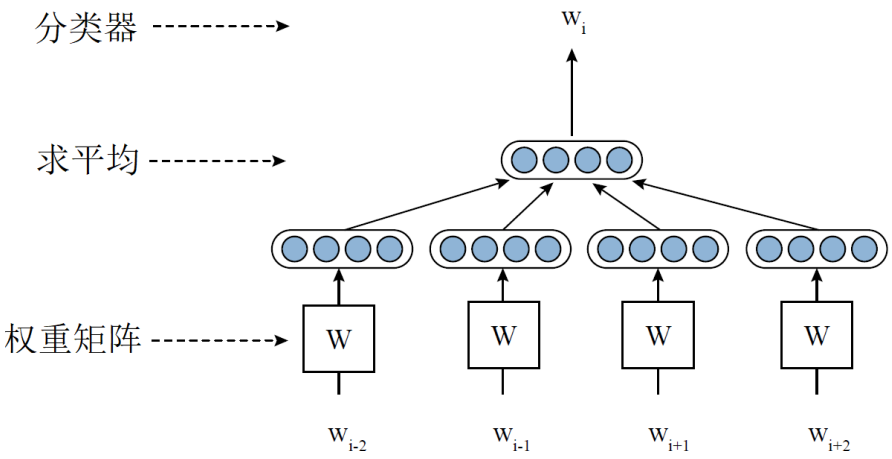


图 2-1 CBOW 模型结构

Figure 2-1 Architecture of Continuous Bag-of-Words Model

（2）Skip-Gram

Skip-Gram 的原理和 CBOW 正好相反，是根据中心词，然后分别计算它前后出现某几个词的概率，计算公式为：

$$P\left(w_j \middle| w_i\right) = Softmax(\boldsymbol{M}w_i)(|j-i| \leq c, j \neq i) \tag{2-2}$$

其中$P\left(w_j \middle| w_i\right)$为给定中心词$w_i$，其窗口大小为 c 的上下文中出现$w_j$的概率。$\boldsymbol{M}$为权重矩阵，意义同上。模型结构如图 2-1 所示：

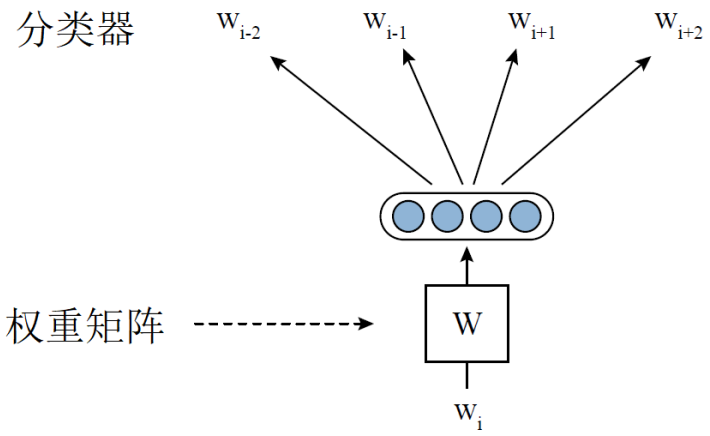


图 2-2 Skip-Gram 模型结构

Figure 2-2 Architecture of Skip-Gram Model

2.2.3 上下文词表征

传统词向量是固定的，不能应词对一词多义的情况。具体来说，相比在词向量矩阵中查表，上下文词表征方法将词和它的上下文一起放进一个深度神经网络，然后得到该词的表示。给定一个序列的 N 个词$(w_1, w_2, \dots, w_N)$，用前向语言模型预测下一个词的概率：

$$P(w_1, w_2, \dots, w_N) = \sum_{k=1}^{N} P(w_k | w_1, w_2, \dots, w_{k-1}) \quad (2\text{-}3)$$

类似的，用反向语言模型预测“下一个”词的概率，其本质是把词序颠倒之后的前向语言模型：

$$P(w_1, w_2, \dots, w_N) = \sum_{k=1}^{N} P(w_k | w_{k+1}, w_{k+2}, \dots, w_N) \quad (2\text{-}4)$$

这里的预测模型可以使用了 BiLSTM，通过前向和后向的语言模型对当前词向量进行微调（fine-tuning），这种能力称之为领域迁移（domain transfer），即使得词向量会动态跟随上下文的场景发生变化，有效的解决了多义词的表征问题。之后的许多预训练模型使用了并行训练效率更高的 Transformer 模型替代 BiLSTM，其基本原理也是基于此双向语言模型。

2.3 机器阅读理解

机器阅读理解（Machine Reading Comprehension，MRC）是一种利用算法使计算机理解文章语义并回答相关问题的技术，其目标是利用人工智能技术，使计算机

具有和人类一样理解文章的能力。

早期的阅读理解模型大多基于检索技术，即根据问题在文章中进行搜索，找到相关的语句作为答案。但是，信息检索主要依赖关键词匹配，而在很多情况下，单纯依靠问题和文章片段的文字匹配找到的答案与问题并不相关。随着深度学习的发展，机器阅读理解进入了神经网络时代。相关技术的进步给模型的效率和质量都带来了很大的提升。机器阅读理解模型的准确率不断提高，在一些数据集上已经达到或超过了人类的平均水平。

基于深度学习的机器阅读理解模型的框架结构一般包含三层：编码层，交互层，输出层。

（1）编码层。

机器阅读理解模型的输入为文章和问题。因此，首先要对这两部分进行数字化编码，变成可以被计算机处理的信息单元。在编码的过程中，模型需要保留原有语句在文章中的语义。因此，每个单词、短语和句子的编码必须建立在理解上下文的基础上。我们把模型中进行编码的模块称为编码层。

（2）交互层

编码层将自然语言映射为数学语言，并没有建立联系性，由于文章和问题之间存在相关性，模型需要建立文章和问题之间的联系。例如，如果问题中出现关键词“河流”，而文章中出现关键词“黄河”，虽然两个词不完全一样，但是其语义编码接近。因此，文章中“黄河”一词以及邻近的语句将成为模型回答问题时的重点关注对象。这可以通过自然语言处理中的注意力机制加以解决。在这个过程中，阅读理解模型将文章和问题的语义结合在一起进行考量，进一步加深模型对于两者各自的理解。我们将这个模块称为交互层。

（3）输出层

经过交互层，模型建立起文章和问题之间的语义联系，就可以预测问题的答案。完成预测功能的模块称为输出层。由于机器阅读理解任务的答案有多种类型，因此输出层的具体形式需要和任务的答案类型相关联。此外，输出层需要确定模型优化时的评估函数和损失函数。

为了降低任务难度，目前很多研究的机器阅读理解都将常识数据排除在外，采用人工构造的比较简单的数据集，回答一些相对简单的问题。常见的机器阅读理解任务可以分为四种类型：完形填空（Cloze-style Test）、多项选择（Multiple Choice）、片段抽取（Span Extraction）、自由回答（Free Answering）。

（1）完形填空

完形填空就是让计算机阅读并理解一篇文章内容后，对机器发出问题，问题往往是抽掉某个词或者实体词的一个句子，而机器回答问题的过程就是将问题句子

中被抽掉的单词或者实体词预测补全出来，一般要求这个被抽掉的词是在文章中出现过的或者根据给定的选项选出正确的答案。具体来说，给定上下文 C，一个词$a \in C$被移除，完形填空任务要求模型使用正确的词或实体进行填空，最大化条件概率$P(a|C-\{a\})$。

（2）多项选择

多项选择与完形填空任务的差异之处在于其对机器发出的问题答案不再局限于词，候选答案通常是一个句子，并且候选答案列表是必须要提供的。给定上下文 C，问题 Q，候选答案列表$A=\{a_1,a_2,\dots,a_n\}$，多项选择任务要求模型从 A 中选择正确的答案a_i，最大化条件概率$P(a_i|C,Q,A)$。

（3）片段抽取

尽管完形填空和多项选择一定程度上具有机器阅读理解的能力，但是这两个任务有一定的局限性。首先，文中的词或实体可能不足以回答问题，需要完整的句子进行回答；其次，在很多情形是没有提供候选答案的。所以片段抽取任务应运而生。给定上下文$C=\{t_1,t_2,\dots,t_n\}$，问题 Q，片段抽取任务要求模型从 C 中抽取连续的子序列$a=\{t_i,t_{i+1},\dots,t_{i+k}\}(1\le i\le i+k\le n)$作为正确答案，最大化条件概率$P(a|C,Q)$。

（4）自由问答

上述三种任务将答案局限于一段上下文，为了回答问题，机器需要在多个上下文中进行推理并总结答案，灵活性较差。自由回答任务是四个任务中最复杂的，也更适合现实的应用场景。给定上下文 C 和问题 Q，在自由回答任务中正确答案可能不是 C 中的一个子序列，即$a \in C$或$a \notin C$，自由回答任务需要预测正确答案 a，最大化条件概率$P(a|C,Q)$。

2.4 开发平台

2.4.1 Colab 在线平台

Google Colab 全名为 Colaboratory，是一个免费的在线 Jupyter Notebook 环境，Colab 的运行原理实际上就是给用户分配一台远程主机。用户只需要打开网页，不需要进行任何设置就可以使用，并且完全在云端运行。借助 Colab，研究者可以编写和执行 Python 代码、保存和共享结果，以及利用强大的 CPU、GPU 和 TPU 计算资源，所有这些都可通过浏览器免费使用。研究者可以用它来提高 Python 技能，也可以使用 Keras、TensorFlow、PyTorch、OpenCV 等流行的深度学习框架开发深度学习应用。

2.4.2 PyTorch 框架

Torch 是采用 Lua 语言为接口的机器学习框架，但是因为Lua语言较为小众，导致 Torch 学习成本高，因此知名度不高。PyTorch 是 Facebook 于 2017 年初在机器学习和科学计算工具 Torch 的基础上，针对 Python 语言发布的一个全新的深度学习框架，该框架基于动态图机制，代码简洁灵活，一经推出便受到了业界的广泛关注和讨论，目前已经成为机器学习从业人员的研发工具。PyTorch 强调灵活性和易用性，使用 Python 语言风格，研究者只需要掌握 Numpy 和基本深度学习概念即可上手，可以很快掌握其语法构建深度学习模型。

PyTorch 提供了一个核心数据结构张量（Tensor），这是一个与 NumPy 数组有很多相似之处的多维数组，但具有 GPU 加速和自动求导的功能，可以加速数学运算。张量包含了数据和元数据(metadata)，元数据用来描述张量的大小、内部数据的类型（整形、浮点型等）和存储的位置(CPU 内存或 GPU 显存)。

就像 Python 用于编程一样，PyTorch 既是深度学习的出色入门框架，也是在专业生产环境中高水平工作的工具。

2.4.3 Jieba 工具包

jieba 是中文“结巴”的拼音。jieba 库是一款优秀的 Python 第三方中文分词库，jieba 支持三种分词模式：精确模式、全模式和搜索引擎模式。精确模式试图将语句最精确的切分，不存在冗余数据，适合做文本分析。全模式将语句中所有可能是词的词语都切分出来，速度很快，但是不能解决歧义，且存在冗余数据。搜索引擎模式在精确模式的基础上，对长词再次进行切分，适合用于搜索引擎分词。

jieba 基于 Trie 树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（Directed acyclic graph)，采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合。而对于未知词，则采用了基于汉字成词能力的隐马尔科夫模型（Hidden Markov Model），使用了维特比（Viterbi）算法。同时开发者可以指定自己自定义的词典，以便包含 jieba 词库里没有的词，虽然 jieba 有新词识别能力，但是自行添加新词可以保证更高的正确率。

除了分词之外，jieba 还有关键词提取、词性标注、词位置检索等功能。

2.5 本章小结

本章主要介绍了工作内容相关的技术背景。首先介绍了自然语言处理的相关内

容，然后介绍本文的背景知识，即中文文本表征的相关概念。作为验证表征质量的基础，2.3 节介绍了基础的四种机器阅读理解任务。2.4 节介绍了本文实验所使用的在线开发平台和深度学习框架以及相关自然语言处理工具包。

3 成语数据集统计分析及任务定义

本章将介绍本文所使用的成语数据集和基线模型的设计以及实验验证。首先对数据集进行统计分析，具体包括数据集的基础统计分析、成语频率分析、相似成语分析，最后介绍不同测试集的难度设置。然后介绍成语机器阅读理解任务的定义，根据编码层、交互层和输出层的三层结构模型设计基线模型架构，并对基线模型在成语数据上的表现进行实验验证。实验结果表明，基线模型的表现还有很大的提升空间。

3.1 成语数据集介绍

本文所使用的是 2019 年清华大学开源数据集 ChID（Chinese Idiom Dataset）[35]1-5。该数据集基于完形填空形式的阅读理解任务，来源于三个领域：新闻、小说和散文，统一为文本数据，每条数据包括文本内容、候选项、问题个数和正确答案，具体信息如表 3-1 所示。

表 3-1 中文成语数据集介绍

Table 3-1 Introduction of Chinese Idiom Dataset

列名	含义
content	文本内容，格式为字符串，其中成语部分用特殊符号#idiom#替换，作为问题项。
candidates	候选答案，格式为列表，列表长度为问题个数，列表中每个元素是长度为 7 的列表，即 7 个成语，其中包含正确答案，3 个相似项，3 个随机项，且三种成语随机分布。
groundTruth	正确答案，格式为列表，列表长度为问题个数，列表中每个元素是一个成语。
realCount	问题个数，格式为短整型，即完形填空需要填空的个数。

数据集示例如表 3-2 所示。其中，文本内容需要填空的位置用特殊符号“#idiom#”替换，作为完形填空的问题标记；候选项中"粉墨登场"为正确答案，"层出不穷"、"纷至沓来"、"风云际会"为相似项，其余为随机项。示例数据中相似项和随机项的比例为 3:3，而根据任务难度，在不同的测试集中相似项和随机项的个数不相同，相似项越多表示该测试集的任务难度越大，后续小节将具体介绍各个测试集。

表 3-2 中文成语数据集示例

Table 3-2 The Example of Chinese Idiom Dataset

列名	数据
content	"2011 年，李晨的两部力作《奋斗》、《建党伟业》将相继上映，电视剧《风车》和《理发师》也将#idiom#，承接 2010 年的辉煌绽放，在新的一年，李晨将展现出更精彩，更惊艳的熟男魅力。"
candidates	[["望洋而叹", "层出不穷", "纷至沓来", "不闻不问", "风云际会", "心广体胖", "粉墨登场"]]
groundTruth	["粉墨登场"]
realCount	1

3.2 数据集统计分析

本小节对成语数据集进行统计分析，主要包括：文本数据的基础统计分析，成语数据的近义词分析，成语词向量语义相似度分析和频率分析。

首先对数据集进行基础统计分析。该数据集覆盖三个领域：新闻、小说和散文，包括 58.1 万条数据，总共 72.9 万个空，平均每条数据下有 1.25 个空，其中最多包含 10 个空，大部分数据空的个数为 1，总计占比约 80%。根据数据领域将数据分为两个部分，新闻和小说作为数据主体领域内（In-domain）部分用于训练数据，散文作为领域外（Out-of-domain）数据用于测试模型的泛化性（generalization ability）。

数据集中的成语来自于《成语大全》[45]。《成语大全》覆盖了超过 2.3 万的四字成语，但大部分为生僻成语，在口语和书面语中都很少使用，因此本数据集只剔选出常用成语。首先，通过已有预训练模型中的成语向量进行筛选，通过检索该成语是否在预训练词向量中存在对应的词向量，过滤掉了约 40%的生僻成语；然后将释义相同的成语视为相同成语进行合并，如“标新创异”、“标新竖异”、“标新领异”、“标新竞异”和的解释均为“同‘标新立异’”，将此类成语合并且使用其所属成语的释义；最后，通过统计成语在数据集中出现的频率，将出现频率小于 20 的成语剔除，最后剩下 3,848 个成语，作为常见成语，具体频率分析如表 3-3 所示，其中后两列为两类文章中成语频率的分布情况。

表 3-3 成语频率分布

Table 3-3 Idiom Frequency Distribution

频率	数量	占比	小说、新闻	散文
[20, 50)	832	21.6%	3.5%	8.2%
[50, 100)	742	19.3%	7.2%	12.0%
[100, 200)	822	21.4%	16.0%	19.7%
[200, 400)	746	19.4%	28.8%	28.7%
[400, 534]	706	18.3%	44.5%	31.4%
[20, 534]	3,848	100%	100%	100%

近义成语的辨别是中文成语一个难点，本数据集的候选答案的设置添加入近义成语以训练模型正确辨别近义成语，同时加入随机成语作为噪声数据。衡量两个词向量之间的语义相似度可以使用余弦距离，但两个成语向量的距离并不一定真实反映其语义相似度。因此，需要手动评估成语向量余弦距离和其语义相似度之间的相关性。首先，将余弦距离区间[0.5, 0.9]等分为 8 个子区间，然后从每个子区间中取样出 200 对成语向量，人工判断两个成语的相似关系，包括三种关系：同义词，近义词和其他，其中同义词指语义相同且在语句中可以互相替换，近义词指语义相近且语句中不可互相替换，其余皆归类为其他。通过四个人进行标注样本后计算得出每个区间中三种类型所占比例，结果如表 3-4 所示，其中 k 为 Fleiss Kappa 系数[46]，用于检验不同标注者结果的一致性。

从表 3-4 中可以看出，余弦距离大于 0.75 的成语对中同义词的比例很大，余弦距离介于 0.65 至 0.80 的则是近义词的比例大。从 Kappa 系数中可以看出，当成语对具有较高（大于 0.85）或较低（小于 0.60）的余弦距离时，标注者倾向于达成一致，而介于[0.65, 0.80]之间则只有中等程度的一致性。基于上述分析，对候选项的设置规则如下：首先，将与正确答案成语余弦距离大于 0.70 的成语排除，避免出现正确答案的同义词；然后，从剩余的成语中（即近义成语）选出余弦距离最大的 10 个成语，并从中随机选择三个作为正确答案的相似项，这三个成语中有可能出现正确答案的同义词，这也一定程度上增加了任务的难度；最后，从剩下的成语中随机抽取三个成语作为随机项，由此便得到了三类一共七个候选项。

为了验证候选项对实验结果的影响，额外设置两个测试用数据集 Ran 和 Sim。Ran 中候选项全部从与正确答案不相似的成语中随机抽取，而 Sim 候选项则从余弦距离前 10 的近义成语中随机抽取。因此普通测试集 Test 中正确答案、相似项、随机项的比例为 1：3：3，Ran 为 1：0：6，Sim 为 1：6：0。Sim 数据集挑战性更强，而 Ran 则更弱。

表 3-4 成语余弦距离与语义相似度的相关性

Table 3-4 The Correlation between Cosine Distance and Semantic Similarity of Idioms

余弦距离	同义词	近义词	其他	k
[0.85, 0.90)	83.2%	16.8%	0.0%	0.642
[0.80, 0.85)	53.6%	42.8%	3.6%	0.447
[0.75, 0.80)	29.2%	53.6%	17.2%	0.485
[0.70, 0.75)	12.0%	57.2%	30.8%	0.496
[0.65, 0.70)	0.4%	52.8%	46.8%	0.466
[0.60, 0.65)	0.0%	34.0%	66.0%	0.528
[0.55, 0.60)	0.0%	10.4%	89.6%	0.657
[0.50, 0.55)	0.0%	6.0%	94.0%	0.787

数据集的具体统计信息如表 3-5 所示。小说和新闻作为主要训练数据，其测试集可以验证模型领域内的迁移能力，散文作为领域外数据验证模型的领域外的迁移能力。数据总共覆盖了 3,848 个常用成语，每个成语出现的平均频率为 189.6。对于领域内数据和领域外数据也有一些差别。首先，散文的平均长度（127）要比小说和新闻的（99）长，而且每个文章中问题平均个数也有差别（1.25 和 1.49）。其次，两类数据的成语频率分布有差异，从表 3-3 中可以看出，在低频区间[20, 50)中，散文数据占比 8.2%，而小说、新闻数据中低频成语占比仅为 3.5%；同样的，在[50, 100)区间中占比分别为 7.2%和 12.0%；而在高频区间[400, 534]中散文数据的占比要比小心、新闻要低。即领域外数据成语的分布特征为：低频率成语占比高，高频率成语占比低，而中频率占比则和领域内数据大致相同。这些统计差异意味着领域外数据的任务难度更高，对模型的挑战性更高。

表 3-5 中文成语数据集统计信息

Table 3-5 Statistics of Chinese Idiom Dataset

	小说、新闻				散文	总计
	训练集	开发集	测试集	总计	总计	
文章个数	520,711	20,000	20,000	560,711	20,096	580,807
平均长度/文章	99	99	99	99	127	100
覆盖成语个数	3,848	3,458	3,502	3,848	3,626	3,848
成语平均频率	168.6	7.2	7.1	181.6	8.3	189.6
问题总个数	648,920	24,822	24,948	698,690	30,023	728,713
平均问题个数/文章	1.25	1.24	1.25	1.25	1.49	1.25
单空比例	80.4%	80.7%	80.8%	80.5%	64.7%	79.9%
多空比例	19.6%	19.3%	19.2%	19.5%	35.3%	20.1%

3.3 成语阅读理解

3.3.1 任务定义

本节将根据上述数据格式定义成语阅读理解任务，并设计基线模型实验验证。

给定文本 C，C 中部分成语$\vec{q} = \{q_1, q_2, \dots, q_n\} \in C$被移除，被移除的位置称为空，候选成语列表$A = \{[a_{11}, a_{12}, \dots, a_{17}], [a_{21}, a_{22}, \dots, a_{27}], \dots, [a_{n1}, a_{n2}, \dots, a_{n7}]\}$，其中 n 为空的个数，从 A 中选择正确的答案$\vec{a} = \{a_1, a_2, \dots, a_n\}(a_i \in [a_{i1}, a_{i2}, \dots, a_{i7}])$，最大化条件概率$P(\vec{a}|C, \vec{q}, A)$。

对于每一个空来说，本质上相当于一个多分类问题，因此本文使用准确率（Precision）评估模型表现：

$$p = \frac{\sum positive}{\sum blank} \tag{3-1}$$

其中$\sum blank$为问题的总个数，$\sum positive$为预测正确的总个数。

3.3.2 基线模型

本节设计成语机器阅读理解的基线模型。模型分为三个部分：编码层、交互层和输出层。

（1）编码层

编码层主要功能是将自然语言进行数字化编码。编码分为两个部分：上下文编码和成语编码。上下文编码与其他自然语言处理任务中正常现代汉语编码相同，涉及到分词、词表映射、向量化。而成语编码为本文研究重点，在第 5 节详细介绍，基线模型中只使用其基本词向量作为成语表征。

首先是词表的获取。使用 Jieba 工具包对每一条文本进行分词，统计整个数据集的词频后，选取频率前 100,000 的词汇，同时将候选项涉及到的成语加入，其余词汇视为未知词汇，使用特殊符号“UNK”（Unknown）表示，由此得到词表。

然后获取词表对应的向量表。本文使用腾讯 AI Lab 开源的大规模高质量中文词向量数据[47]获取需要的词向量，该数据包含 800 多万中文词汇，语料来自腾讯新闻、天天快报的新闻语料，互联网网页和小说语料，在覆盖率、新鲜度及准确性上相比其他公开数据集大幅提高，其中每个词对应一个 200 维的向量。从这 800 多万向量中选出本文需要的 100,000 个词向量，同时也选出候选项涉及到的成语向量，而对于“UNK”则按照范围（-0.1，0.1）的均匀分布（Uniform distribution），将其随机初始化，由此得到词向量表。

对于一条数据，上下文编码处理步骤如下：

1）使用 Jieba 对文本进行分词，得到词汇列表，此时列表中元素为词；

2）将词汇列表映射为词汇索引列表，即将列表中的每个词映射为其在词表中的索引，此时列表中元素的索引；

3）将词汇索引表映射为词向量列表，即将列表中的每个索引映射为对应向量表中的词向量，得到上下文最终的向量形式。

词向量列表本质上相当于一个矩阵，对于长度为 L 的上下文，其词向量列表相当于 L 行 200 列的矩阵，矩阵每一行为一个 200 维的词向量。

对成语进行编码和上述步骤相似，只是少了第一步分词。

（2）交互层和输出层

交互层主要功能是建立向量之间的联系以模拟语义关系，输出层则根据交互层的状态预测正确答案。本文使用三种传统阅读理解模型：BiLSTM[48]、AR[49]和 SAR[50]。

1）BiLSTM

BiLSTM 全称为 Bi-directional Long Short-Term Memory，即双向长短期记忆网络，由前向 LSTM 和后向 LSTM 组合而成，分别用于对上文和下文进行建模。LSTM 模型在训练过程中可以学到记忆哪些信息和遗忘哪些信息，从而可以捕捉到较长距离的词与词的依赖关系。具体计算公式如式 3-2 至 3-5 所示。

$$\overrightarrow{\boldsymbol{h}_b} = \overrightarrow{LSTM}(w_{1:b}) \tag{3-2}$$

$$\overleftarrow{\boldsymbol{h}_b} = \overleftarrow{LSTM}(w_{b:|P|}) \tag{3-3}$$

$$\boldsymbol{h}_b = \overrightarrow{\boldsymbol{h}_b} \,||\, \overleftarrow{\boldsymbol{h}_b} \tag{3-4}$$

$$\alpha_i = softmax_i(\boldsymbol{h}_b^T \boldsymbol{a}_i) \tag{3-5}$$

式 3-4 使用 BiLSTM 对文本进行建模，获取空的位置（以下简称为空）的隐藏层状态（hidden state）$\boldsymbol{h}_b$（b 表示 blank），$\boldsymbol{h}_b$由前向结果$\overrightarrow{\boldsymbol{h}_b}$和后向结果$\overleftarrow{\boldsymbol{h}_b}$拼接而成，其中|P|为文本长度，$w_{1:b}$和$w_{b:|P|}$分别表示空的上文词向量和下文词向量，||表示将两个向量拼接。如式 3-5 所示，获取每个空位置的隐藏层状态后，与各自的候选答案计算得出每个候选答案的得分α_i，其中$\boldsymbol{a_i}$表示第 i 个候选成语的表征，基线模型中使用成语的词向量作为其表征，选择得分最高的候选项作为预测答案。

2）AR

AR 全称为 Attentive Reader，是在 BiLSTM 基础上，使用注意力机制对模型的阅读理解能力进行增强。具体计算公式如式 3-6 至 3-10 所示，其中W_{hm}、W_{bm}、w_{ms}、W_{rg}、W_{bg}为神经网络的超参数，tanh 为双曲正切函数。

$$\boldsymbol{m}_t = \tanh\left(W_{hm}\boldsymbol{h}_t + W_{bm}\boldsymbol{h}_b\right) \tag{3-6}$$

$$s_t = softmax_t(w_{ms}^T \boldsymbol{m}_t) \tag{3-7}$$

$$\boldsymbol{r} = \sum_{t=1}^{|P|} s_t \boldsymbol{h}_t \tag{3-8}$$

$$\boldsymbol{g} = \tanh\left(W_{rg}\boldsymbol{r} + W_{bg}\boldsymbol{h}_b\right) \tag{3-9}$$

$$\alpha_i = softmax_i(\boldsymbol{g}^T \boldsymbol{a}_i) \tag{3-10}$$

式 3-6 至 3-7 为注意力机制的计算过程。式 3-6、3-7 使用空的隐藏层状态$\boldsymbol{h}_b$与其他词的隐藏层状态$\boldsymbol{h}_t$进行交互，计算得到其他词与该空的关联性大小s_t，然后式 3-8 将s_t作为权重进行加权求和得到“注意力”向量$\boldsymbol{r}$，$\boldsymbol{r}$中蕴含了整个文本的语义，且由于加入了权重s_t，使得语义的“注意力”更多的关注于与空有关联的词汇。式 3-9、3-10 使用“注意力”向量$\boldsymbol{r}$、空的状态$\boldsymbol{h}_b$与候选项$\boldsymbol{a}_i$计算得出每个候选答案的得分α_i。

3）SAR

SAR 全称为 Stanford Attentive Reader，AR 注意力机制的权重计算使用了双曲正切函数 tanh，而 SAR 则使用了双线性函数，如式 3-11 至 3-13 所示，其中W_s为神经网络超参数。

$$s_t = softmax_t(\boldsymbol{h}_b^T W_s \boldsymbol{h}_t) \tag{3-11}$$

$$\boldsymbol{o} = \sum_{t=1}^{|P|} s_t \boldsymbol{h}_t \tag{3-12}$$

$$\alpha_i = softmax_i(\boldsymbol{o}^T \boldsymbol{a}_i) \tag{3-13}$$

如式 3-11 所示，SAR 模型中注意力机制权重的计算使用了双线性项W_s，本质上相当于$\boldsymbol{h}_t$和$\boldsymbol{h}_b$点乘的线性变换，而 3-7 中 AR 模型$\boldsymbol{h}_t$和$\boldsymbol{h}_b$的交互方式则采用了二者线性变换后相加，所以 AR 和 SAR 本质上区别在于注意力机制中权重计算分别采用了加法和点乘。

模型的损失函数为交叉熵损失函数，如式 3-14 所示，其中y_j根据预测正确与否为 0 或 1，α_i为每个答案的概率，求和公式中的 7 为候选答案个数。

$$Loss = \sum_{j=1}^{7} y_j \log\left(\alpha_i\right) \tag{3-14}$$

加入注意力机制 AR、SAR 的模型整体架构如图 3-1 所示，其中 BiLSTM 模型与 AR、SAR 模型的区别在于 BiLSTM 没有中间注意力机制的计算过程，基线模型的编码层仅使用词向量编码。

3.3.3 实验结果

本文实验参数设置如下：本文实验均基于 PyTorch 框架实现，文本预处理使用 Jieba 分词软件包，神经网络相关参数中 epoch 设置为 10，batch size 设置为 32，词表征层的 dropout 概率为 0.5，优化器使用 Adam[51]，初始学习速率为 0.001，BiLSTM 模型梯度修剪（clip）阈值为 5.0。每训练 1000 次 batch 在 Dev 测试集上进行测试，当在 Dev 上测试结果连续两次没有提升时，减小训练速率，减小幅度为 0.95，当

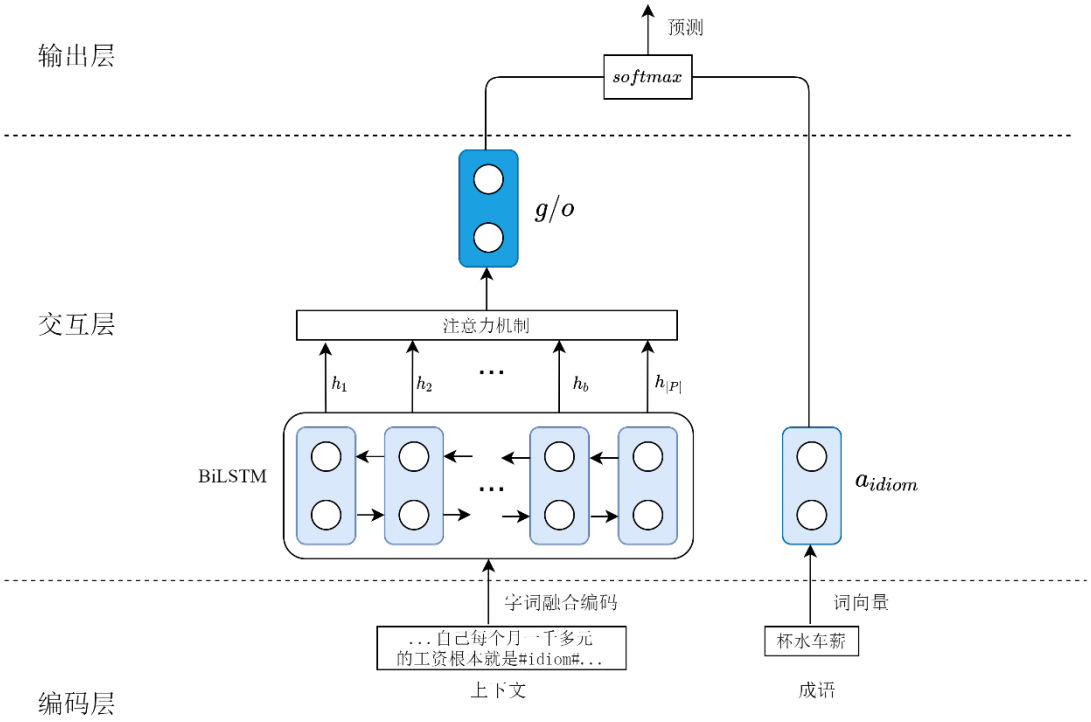


图 3-1 模型整体框架

Figure 3-1 Architecture Overview

在 Dev 上测试结果在一轮（epoch）中都没有提升时，提前停止训练。

实验结果如表 3-6 所示，其中第二列第三列含义式基线模型中上下文和成语的表征均使用词向量，Human 结果为 3 位清华大学大一或大二的学生在随机抽取的 800 道题上的测试结果[35]7，k 为 Kappa 系数，用于衡量 3 位测试者结果的一致性。

表 3-6 基线模型表现

Table 3-6 Performance of Baseline Models

	Dev	Test	Ran	Sim	Out
Human	-	87.1	97.6	82.2	86.2
k	-	0.794	0.953	0.791	0.769
BiLSTM	68.8	68.7	77.9	64.3	59.5
AR	69.8	69.6	79.2	65.7	60.9
SAR	68.9	68.7	78.2	64.4	60.1

从纵向结果中可以看出，AR 模型在各个测试集中表现最好，原因在于 AR 模型使用了空的隐藏层状态$\boldsymbol{h}_b$与上下文状态$\boldsymbol{h}_t$进行交互（式 3-6），实现了根据空“细读”上下文，然后再使用空的状态与“注意力”向量$\boldsymbol{r}$对候选项进行评分（式 3-9），

通过这种方式选择性的细读上下文，并且空的状态使用了两次，模型性能较好。SAR 相比 BiLSTM 只有较小的性能提升，也说明了注意力机制中权重计算函数的选择对其性能表现影响较大。

然后对结果进行横向对比。首先，可以看出各个模型在测试集上的表现由高到低依次为：Ran>Test & Dev> Sim > Out，表现与候选答案中相似项所占比例成反比，即相似项越多任务难度越大；其次，从 Kappa 系数结果可以看出，3 位测试者在 Ran 测试集上的一致性最高（0.953），且准确率（97.6%）远高于其余测试集，这是因为 Ran 中没有相似项的干扰，任务难度较低，对于测试者来说很容易选出正确答案。而出现相似项的测试集中，测试者的准确率也随着相似项的增加而降低，且三位测试者答案的一致性也较 Ran 差，这说明相似成语的辨别对于人类来说也是一项艰难的挑战。

3.5 本章小结

本章主要介绍了成语数据集的分析，然后实验验证了传统模型在该数据集上的表现。成语数据集分析主要包括：数据展示及每一列含义，数据集覆盖的领域，覆盖成语的设置，成语频率分析，相似成语语义分析以及领域内和领域外数据任务难度分析。接着 3.3 节对成语阅读理解任务进行了定义，实验设计了基线模型，并对实验结果进行横向和纵向分析。

4 基于字词融合的上下文表征

本章针对阅读理解中问题的上下文，提出一种基于字词融合的神经网络模型。为解决字向量和词向量数目不一致的问题，本章提出两种字词对齐模型；为了探究有效的字词信息交互的方式，本章还提出了三种字词融合方法；最后，为了验证成语的基于语义特性，本章设置了相关的对照试验。下面具体介绍本章的基本思想和模型实现细节。

4.1 基本思想

词级别表征主要面临两个问题：数据稀疏（Data Sparsity）和未知词（Out-of-Vocabulary Words）。

数据稀疏指的是一些词出现的频率较低，导致得到的词向量质量较差。在词级别的表征中，词的频率分布不均匀，很多词的频率很低，即数据稀疏，而对于模型来说，要学习一个词的语义信息，该词的频率需要达到一定量才能获得质量较高的词向量。因此，在词级别模型中，神经网络并没有充分学习到很多低频词的语义信息。

中文的词由汉字组合而成，词的长度不固定，且新词不断涌现，这就导致了词表中不可能收录所有词汇。我们把文本分词结果中不存在于词表中的词称为未知词，使用特殊记号 UNK（Unknown）来表示这些词。而对于未知词的定义，通常做法是对未知词设置一个词频门限(Frequency threshold)，出现次数低于该门限的词就称为未知词。对于未知词来说，因为都将其归类为了 UNK，使用一个随机初始化的词向量来表示 UNK，即后面的所有未知词共享了一个词向量，模型就比较困难去学习到他们的语义信息。虽然可以设置一个比较低的门限，但是这样会导致数据集中出现很多低频词，又会产生数据稀疏问题。

现代汉语中常用汉字只有几千个，生僻字在一般的文章中极少使用，因此字向量的获取过程中不会遇到数据稀疏问题，且未知字的情况也极少发生。字的粒度比词的粒度小，可以学习到更多细粒度的语法依存关系，字信息可以更细粒度的方式对词信息进行补充，并解决分词带来的问题。尽管字级别模型听起来很有潜力，但它们确实也违反直觉，因为大部分字只有组成词才具有语义含义，单独的字符则没有。因此本文将字词信息进行融合，对文本编码层进行改进。

字向量的获取方法与 3.3.2 节中词向量的获取方法类似：统计数据集中字个频率后，分析仅出现一次的汉字，发现均是生僻字，如“骾”、“蓐”、“爨”等，将出现频率只有一次的删除；然后删除特殊符号，如“@”。经过清洗后字表大小为 6872，

从预训练向量中提取出本文需要的 6872 个字向量，对于不在字表中的符号和字，使用特殊符号 UNK_CHAR（Unknown character）代替，按照范围（-0.1，0.1）的均匀分布（Uniform distribution）随机初始化 UNK_CHAR，由此得到字向量表。

本章提出字词融合的思想，用以解决单一词向量或字向量无法有效表针的问题，同时比较不同融合方法，比较不同融合方法的有效性。同时，也对成语字词融合进行实验分析，侧面论证成语的语义特性。

4.2 模型结构

本节将具体介绍本文提出的字词融合模型的模型结构及其构建过程。

首先需要对字词向量进行对齐操作。对文本进行字级别编码得到的向量列表的长度比词级别划分的长，要融合二者的信息，就要进行字词的对齐操作，便于字词融合。本文采用两类对齐方法，一种是对组成一个词的字向量进行操作使其与词向量对齐，另一种是对词向量进行操作使其与字向量对齐。

对齐之后本文采用三种融合方法对字词信息交互进行建模，分别是元素乘（element-wise multiplication）、元素加（element-wise summation）和拼接（concatenation,）。其中，元素乘/加操作是两个向量对应维度的元素相乘/加，得到的结果维度不变；而拼接操作是两个向量首尾拼接，得到的结果维度翻倍。

4.2.1 字向词对齐

字向词的操作目标是将多个字向量通过神经网络操作，学得其字与字之间语义关系，得到一个向量，然后与词向量进行融合。本文使用双向门限循环网络（Bidirectional Gated Recurrent Unit, BiGRU）对字向量进行处理，GRU 可以看成是 LSTM 的变种，GRU 把 LSTM 中的遗忘门和输入们用更新门来替代，且在计算新信息的方法和 LSTM 有所不同，具体公式如式 4-1 至 4-4 所示：

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{4-1}$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{4-2}$$

$$\widetilde{h_t} = \tanh(W_h x_t + U_h(rt \odot h_{t-1}) + b_h) \tag{4-3}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot h_t \tag{4-4}$$

其中r_t为重置门（reset gate），z_t为更新门（update gate），$\widetilde{h_t}$为隐藏层状态，σ为 sigmoid 激活函数，W 和 b 为神经网络超参数。重置门决定了如何将新的输入信息与之前的记忆信息相结合（式 4-3），更新门则定义了之前记忆信息保存到当前时间步的量（式 4-4）。BiGRU 的和 BiLSTM 的思路一样，$\overleftrightarrow{h_t}$为其前向 GRU 输出$\overrightarrow{h_t}$和后向 GRU 结果$\overleftarrow{h_t}$拼接而成：$\overleftrightarrow{h_t} = \overrightarrow{h_t}||\overleftarrow{h_t}$。然后将 BiGRU 的输出结果，即将每

个方向最后位置的隐状态拼接，经过一个全连接层，得到最终字级别的表征，如式4-5所示，其中W和b为全连接层神经网络超参数：

$$CE(w) = W\overleftrightarrow{h_t} + b \tag{4-5}$$

$CE(w)$（Character Embedding）即为词 w 的字表征，将其与该词的词表征$WE(w)$(Word Embedding)融合后得到字词融合表征$ME(w)$（Mixed Embedding），如式4-6所示，其中•表示元素乘、元素加和拼接三种融合方法：

$$ME(w) = WE(w) \bullet CE(w) \tag{4-6}$$

以词“大学生”为例，字向词对齐模型框架如图4-1所示。其中右侧为词表征的获取，直接在词向量表中查询得到其词向量作为其词表征$WE(w)$；左侧为字表征的获取，绿色部分为字向量相关的处理模块，首先在字向量表中查表得到三个字的词向量，然后通过BiGRU层和全连接层将三个字向量与词表征对齐，得到字表征$CE(w)$；最后将二者融合得到字词融合表征$ME(w)$。

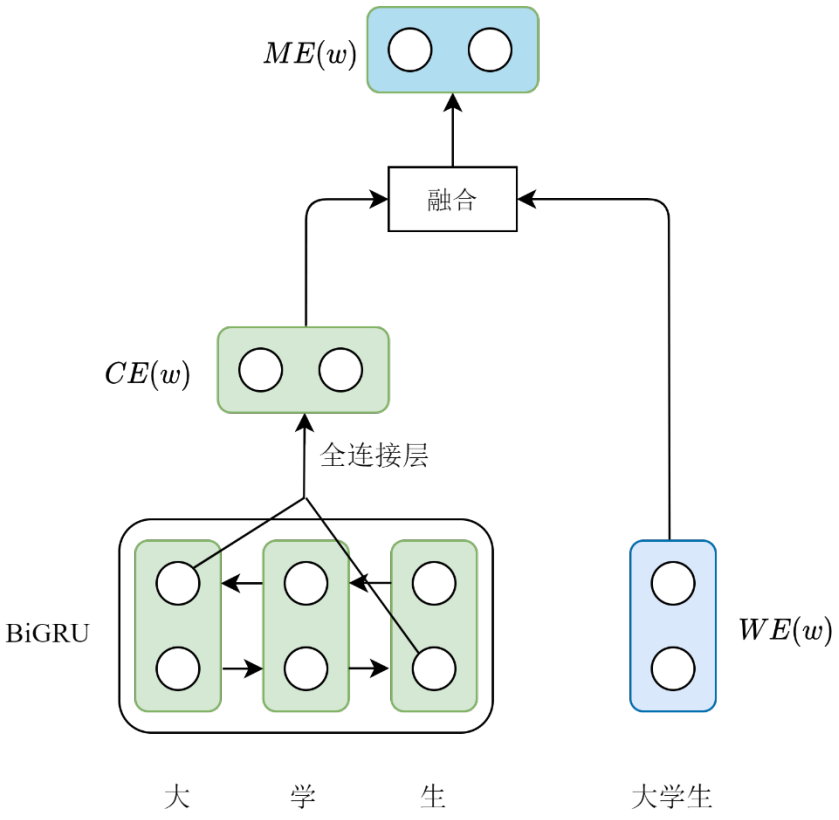


图4-1 字向词对齐模型结构

Figure 4-1 Architecture of Character-to-word Alignment Model

4.2.2 词向字对齐

词向字对齐的目标是将一个词向量与多个字向量交互，该操作需要保证每个字与同一个词向量交互，相当于使用字信息对词信息的语义进行微调（Fine-tuning），同时对于未知词 UNK 来说，其字信息可以有效弥补其词信息的损失。本文采取直接将词向量进行复制的方法，然后和字向量进行对齐。

同样以词“大学生”为例，词向字对齐模型框架如图 4-2 所示。首先根据词的长度将词向量复制 3 份得到词表征$WE(w)$，然后与每个字向量进行融合操作得到融合表征$ME(w)$。

如图 4-1 和 4-2 所示，字向词对齐是字向量的“收缩”，而词向字对齐则是词向量的“扩展”，所以对于同一文本来说，经过编码层后二者得到的向量列表长度会相差很多，由于中文二字词居多，所以长度差距约为 2 倍多，长度的差异也会对阅读理解模型的最终结果产生影响。

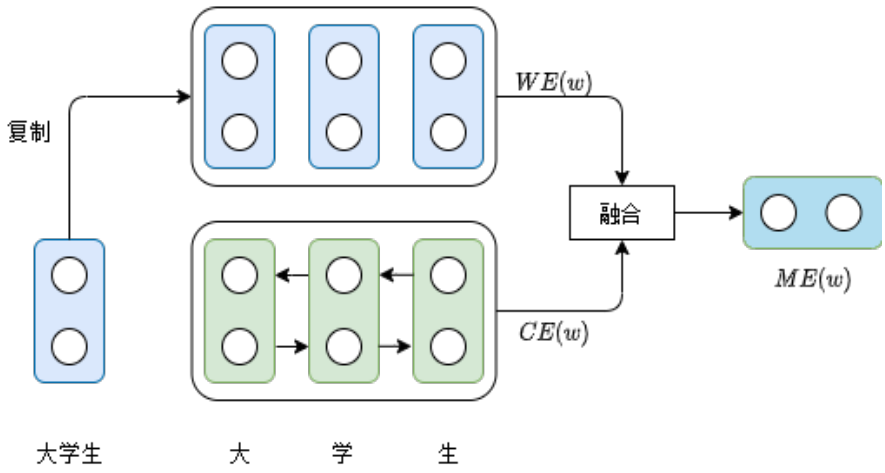


图 4-2 词向字对齐模型结构

Figure 4-2 Architecture of Word-to-character Alignment Model

4.3 实验验证

本节对 2.4 中提出的模型进行实验验证，该实验是针对上下文文本编码的，因

此首先在 4.3.1 节中成语编码不变，通过对两种字词对齐方式和三种字词融合方式的结果进行对比分析，找出最佳的上下文文本表征组合方式。然后 4.3.2 节使用字词融合的方法对成语进行编码，分析实验结果，以验证成语的语义特性。

4.3.1 字词融合网络模型

字词融合模型主要针对现代汉语形式的上下文文本，实验结果如表 4-1 所示，其中成语表征采用 3.3.2 小节中所述的词向量。从结果中可以得出如下结论：

（1）从总体的纵向结果来看：字向词对齐方式和元素乘的组合表现最好，词向字对齐方式和拼接的组合表现最差，这说明字信息的引入确实能够丰富词信息，提升模型性能。但是，字信息需要以有效的方式与词信息交互，从词向字对齐方式和拼接的组合结果中可以看出，其整体性能相对于基线模型甚至有所下降。而其他组合方式相比于基线模型，性能最高提升了 6.3%（AR 模型在 dev 数据集上从 69.8 提升到 74.2）。

（2）从对齐方式分析：在融合方式和阅读理解模型一致的情况下，字向词对齐的整体性能要优于词向字对齐。这是由中文的语法特点决定的，即中文中词往往是信息载体的最小单位，字往往是没有明确含义的，同时单个字的含义过多不便于机器识别，或者说不容易用单个向量表示其含义，而和前后的字构成固定词组后，歧义量就被缩减了，即不确定程度降低了，因此词的蕴含的语义信息确定性强，所以以词向量为基础的阅读理解模型性能要更好。此外，词向字对齐方法得到的向量列表长度更长，约为字向词对齐方法的 2 到 3 倍，这就要求阅读理解模型能够捕获更长距离的依赖关系。因此，字向词对齐方法的优势在于不改变编码后向量列表的长度，避免了因为编码结果长度差异带来的性能损失，有效地实现了字信息对词信息的补充或微调。

（3）从融合方式分析：在对齐方式和阅读理解模型一致的情况下，元素乘的整体性能最优，元素加次之，拼接的方法表现最差。元素乘的优越性在于，其可以对词向量和字向量每一维度的交互进行建模，并消除或减小两种向量在向量空间分布的差异，而且元素乘类似于注意力机制，即对词向量的维度层面运用注意力机制，而每一维度的权重则由字信息决定。元素加也是对向量的每一维度进行交互，但是由于求和运算过于简单，而无法有效的对字词交互进行建模。拼接操作的表现最差，这是因为字词向量的每一维度并没有进行交互，而且编码后向量的维度翻倍，造成了严重的过拟合问题，导致性能下降。

表 4-1 基于字词融合的模型表现

Table 4-1 Performance of Mixed Embedding of Words and Characters

	Dev	Test	Ran	Sim	Out
基线模型					
BiLSTM	68.8	68.7	77.9	64.3	59.5
AR	69.8	69.6	79.2	65.7	60.9
SAR	68.9	68.7	78.2	64.4	60.1
字向词对齐+元素乘					
BiLSTM	72.9	72.8	80.6	67.1	63.3
AR	**74.2**	**73.9**	**82.2**	**68.6**	**64.1**
SAR	73.0	72.6	80.5	67.3	62.9
字向词对齐+元素加					
BiLSTM	70.8	70.8	78.9	65.7	61.8
AR	71.9	72.1	80.5	66.6	62.7
SAR	71.1	70.7	79.4	65.5	62.0
字向词对齐+拼接					
BiLSTM	69.5	69.1	78.4	64.2	60.5
AR	69.3	70.3	79.7	66.1	61.2
SAR	69.7	68.4	78.6	64.6	60.8
词向字对齐+元素乘					
BiLSTM	71.6	72.0	79.9	66.7	62.8
AR	72.5	72.6	80.8	67.8	63.9
SAR	71.8	72.1	80.2	66.5	62.8
词向字对齐+元素加					
BiLSTM	70.7	71.1	79.6	65.1	61.9
AR	71.6	71.9	80.3	65.9	62.5
SAR	70.6	71.2	80.1	65.3	61.6
词向字对齐+拼接					
BiLSTM	68.3	68.5	78.2	64.4	59.9
AR	69.5	69.8	79.1	66.1	60.7
SAR	68.4	68.8	77.6	64.2	59.5

4.3.2 成语基本语义特性

在 1.3 节中，本文介绍了成语的基本特性之一：非语义合成性和意义整体性，在本节通过具体的实验验证成语的这一特性。在上一节的实验中，成语编码使用了成语的词向量，没有涉及到成语的字信息，同时字词融合编码中字向词对齐和元素乘的组合方式性能表现最好。因此本节使用该组合对成语进行编码，上下文文本编码则使用了词向量作为对照。实验结果如表 4-2 所示，其中“字词融合”指的是使

用字向词对齐和元素乘的组合方式进行编码。从结果中可以看出：

（1）从成语表征的角度分析：当将成语编码由词向量变为字词融合时，模型表现变得非常差，准确率下降了约 10%，这表明成语的字信息的引入反而使其词向量的表征能力下降了，这也是人类正确理解中文成语的难点之一：望文生义，即“不了解某一词句的确切涵义或来源缘由，光从字面上去牵强附会，做出不确切的解释”。该结果也从侧面印证了成语的非语义合成性和意义整体性。

（2）从阅读理解模型角度分析：当引入成语的字信息后，AR 和 SAR 的性能反而不如朴素的 BiLSTM，这说明成语阅读理解模型不仅要有合适的模型结构，也需要对成语进行有效的表征，本文将在在下一章中针对成语表征进行增强改进。

表 4-2 成语非语义合成性验证实验

Table 4-2 Experiment on the Non-compositionality of Idioms

	Dev	Test	Ran	Sim	Out
基线模型					
BiLSTM	68.8	68.7	77.9	64.3	59.5
AR	69.8	69.6	79.2	65.7	60.9
SAR	68.9	68.7	78.2	64.4	60.1
成语字词融合编码					
BiLSTM	63.1	63.4	74.4	61.0	55.8
AR	62.5	61.5	72.6	59.9	54.4
SAR	62.9	61.8	72.5	60.2	55.6

4.4 本章小结

本章提出了针对现代汉语的字词融合编码，来改进对上下文文本表征，同时也对成语的非语义合成性和意义整体性进行实验验证，具体工作如下：

（1）针对文本按照词切分和按照字切分结果长度不一致的问题，本文提出两种对齐方式：字向词对齐和词向字对齐，同时针对字词交互的方式，本文又提出三种字词融合的方法：元素乘、元素加和拼接。实验结果表明字向词对齐和元素乘的组合方式能够有效的对字词信息进行整合，性能最高提升了 6.3%。

（2）为了验证成语的基本语义特性，本文又设置了对照实习，对成语进行字词融合的编码，实验结果表明成语字信息的引入降低了模型的整体性能，证实了成语的非语义合成性。

5 基于释义增强的多粒度成语表征

上一章阐述了成语的字信息无法对成语的词信息进行信息补充，反而降低了性能。为了提升成语表征的质量，本章引入了成语的现代汉语解释，提出成语释义增强（Definition-augmented Embedding for Chinese Idiom）神经网络模型。成语释义增强模型主要挑战在于释义信息中不同成分的筛选以及和词信息的结合。为了解决这一问题，本文基于注意力机制，实现了在词信息和释义信息有效结合的同时，对释义信息中不同成分进行了有效地筛选。下面将具体介绍模型的原理，通过实验验证其效果并进行可视化分析。

5.1 基本思想

成语的意义具有较强的整体性和抽象延伸性，如“洛阳纸贵”的释义为“原指洛阳之纸，一时求多于供，货缺而贵。比喻文章写得好，风行一时。”，从具体事务中延伸出抽象的含义是部分中文成语的一大特点。

对于人来说，正确理解这些成语需要了解其背后额外的复杂背景知识。而对于机器来说，仅靠成语的词向量只能对其语义进行粗粒度的表征，无法有效区分相似成语之间的细微差别，即词级别的表征缺乏细粒度信息，因此需要引入额外的背景知识，即成语的释义，对成语的表征进行细粒度级别的信息增强。

成语释义信息的引入主要面临两个难点：

（1）成语释义的数据收集与清洗。成语释义中包含了许多文言文以解释其出处，如“杯水车薪：意思是用一杯水去救一车着了火的柴草，比喻力量太小，解决不了问题。出自《孟子·告子上》：今之为仁者，犹以一杯水救一车薪之火也。”，即成语释义的内容多为现代汉语与古汉语混杂的形式，因此需要对成语释义数据集进行清洗以降低噪声数据的影响。

（2）多粒度信息的有效结合。成语释义中的不同部分对其语义的贡献大小不同，如上述“杯水车薪”的释义中贡献较大的部分为“力量太小，解决不了问题”，因此如何对释义中的不同部分进行有效的筛选，以及如何将筛选后的信息与其粗粒度的词信息进行有效的整合，是释义信息引入的另一个困难。

对于释义数据的收集，首先从《成语大全》[36]中索引出成语的原始释义，然后进行人工清洗，清洗原则为只保留现代汉语的解释，将成语的出处的文言文部分删除，释义清洗的一个例子如表 5-1 所示。而对于释义信息的筛选，本文基于注意力机制，结合成语的词信息，实现有效的筛选。

表 5-1 成语释义数据清洗示例

Table 5-1 Example of Idiom Definition Data Cleaning

成语	一心一意
原始释义	《三国志·魏志·杜恕传》“免为庶人，徙章武郡，是岁嘉平 元年”裴松之注引《杜氏新书》：“故推一心，任一意，直而行之耳。”后因以“一心一意”谓同心同意；或专心专意，毫无他念。
清洗后释义	同心同意；或专心专意，毫无他念。

5.2 模型结构

本节介绍成语多粒度信息的整合。成语释义信息的基本表征的获取方法采用 4.3 节提出的字词融合方法，即字向词对齐和元素乘的组合方式。

5.2.1 成语释义增强模型

上一节分析中提到，成语释义中不同成分对其语义的贡献程度不同，即需要在细粒度级别筛选出贡献程度较大的部分，因此使用 BiLSTM 和注意力机制动态地调整成语释义的权重，其中查询（Query）向量使用成语词向量，即成语词向量分别与释义中的不同部分进行交互，赋予经过 BiLSTM 后各个部分的隐状态以不同的权重，最后加权求和，具体计算公式如 5-1 至 5-6 所示。

$$\overrightarrow{\boldsymbol{h}_t} = \overrightarrow{LSTM}(\boldsymbol{d}_{1:t}) \tag{5-1}$$

$$\overleftarrow{\boldsymbol{h}_t} = \overleftarrow{LSTM}(\boldsymbol{d}_{t:|D|}) \tag{5-2}$$

$$\boldsymbol{h}_t = \overrightarrow{\boldsymbol{h}_t} \,||\, \overleftarrow{\boldsymbol{h}_t} \tag{5-3}$$

$$\boldsymbol{e_t} = \boldsymbol{w}_{idiom} W_d \boldsymbol{h_t} \tag{5-4}$$

$$s_t = softmax(\boldsymbol{e_t}) \tag{5-5}$$

$$\boldsymbol{a}_{idiom} = \sum s_t \boldsymbol{h_t} \tag{5-6}$$

其中$\boldsymbol{d}$为成语释义的词向量列表（这里的词向量列表指的是经过第 4 节介绍的方法编码后，得到的向量列表，下同），$\boldsymbol{d}_{1:t}$和$\boldsymbol{d}_{t:|D|}$分别为位置 t 的词的上文和下文，|D|为释义词向量列表长度，经过前向（式 5-1）和后向（式 5-2）LSTM 后，式 5-3 将二者隐状态拼接得到位置 t 的隐状态向量$\boldsymbol{h}_t$；式 5-4 和式 5-5 为注意力机制的计算过程，其中查询（Query）向量为成语的词向量$\boldsymbol{w}_{idiom}$，分别与每个位置的隐状态向量$\boldsymbol{h_t}$进行交互，W_d为神经网络超参数，用于对成语词向量和成语释义的相关性进行建模；最后式 5-6 将成语释义的词向量列表进行加权求和得到成语的

表征$\boldsymbol{a}_{idiom}$，这里的$\boldsymbol{a}_{idiom}$是式 3-5、3-10 和 3-13 中的$\boldsymbol{a}_i$，用于与阅读理解模型交互，预测正确答案。

以成语“杯水车薪”为例，成语释义增强模型流程如图 5-1 所示。图中左侧为成语的词向量，右侧为对成语释义进行字词融合编码后，将其送入 BiLSTM 对释义中的语义关系进行建模，然后经过一层注意力机制，该注意力机制将成语的词信息和释义信息进行了有效地整合，最后加权求和，得到最终的成语表征$\boldsymbol{a}_{idiom}$。

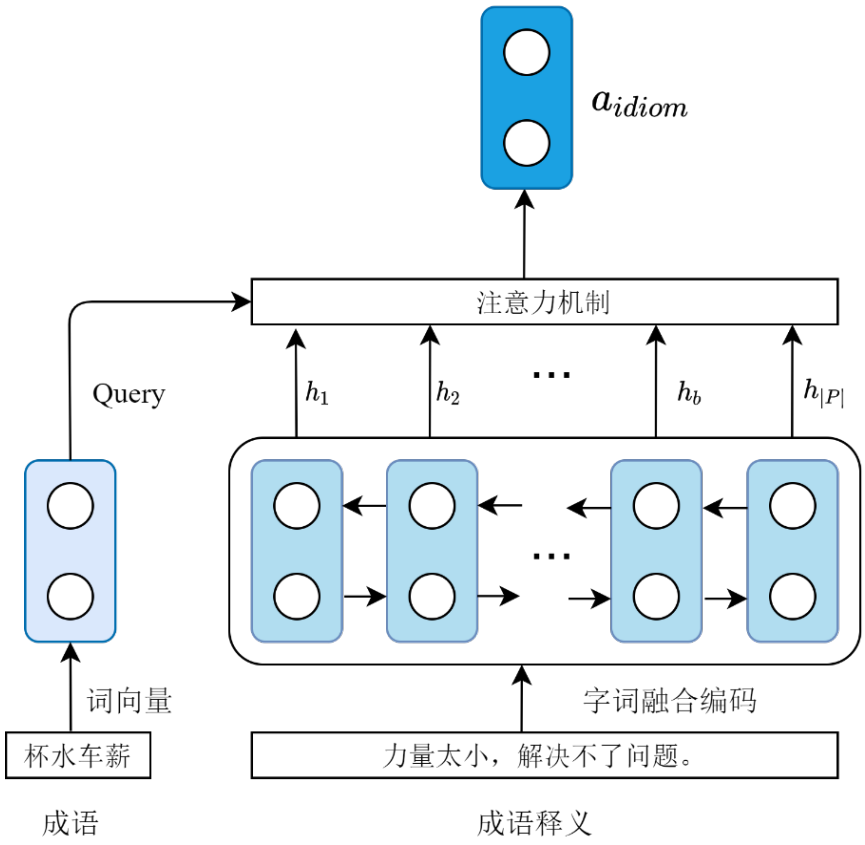


图 5-1 基于注意力机制的成语释义增强模型结构

Figure 5-1 Architecture of Definition-augmented Model Based on Attention Mechanism

阅读理解模型整体结构如图 5-2 所示。该模型结构相较于基线模型结构（图 3-1），主要区别在于对编码层进行了改进，具体包括两个部分：

（1）上下文的字词融合编码。图中编码层左侧部分为上下文编码，采用 4.2 节提出的字向词对齐和元素乘的组合方式。

（2）成语的释义增强编码。图中编码层右侧部分为成语释义增强编码，即图 5-1 所述模型。

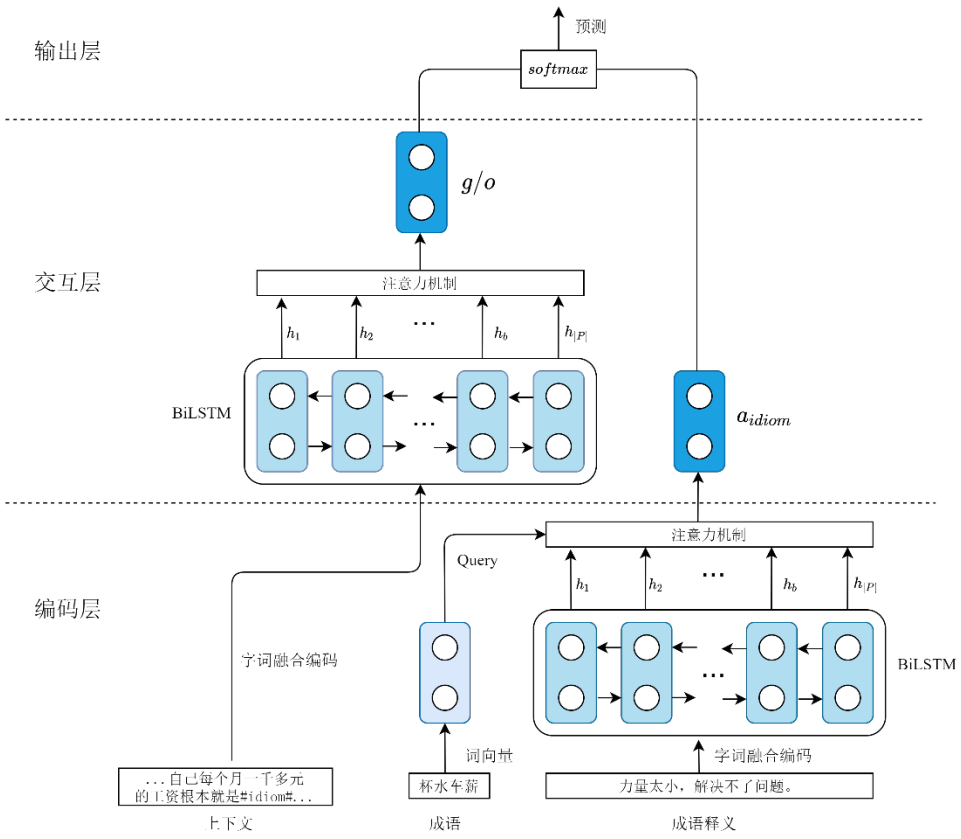


图 5-2 成语释义增强模型整体架构

Figure 5-2 Architecture Overview of Definition-augmented Embedding for Idioms

5.2.2 对照实验

为了验证释义中不同成分的贡献不同的猜想，本节设置相关的对照实验。如上分析，式 5-4 和 5-5 为注意力机制的计算过程，相当于计算释义中不同成分的权重，因此对照组的设置没有注意力机制的计算过程。同时对照组也可以验证仅使用成语释义对成语进行表征的效果。

对照组采用两种常用方法得到最终的释义表征：

（1）取前向 LSTM 和后向 LSTM 的最后一个位置隐状态向量进行拼接，每个方向的最后一个隐状态中蕴含了其前方所有隐状态的信息，并通过 LSTM 的“门”机制动态地在信息传递过程中判断历史信息保留程度，该组对照实验用于模拟忽略成语的词信息而仅考虑细粒度信息，如式 5-7 所示，模型流程图如图 5-3 所示。

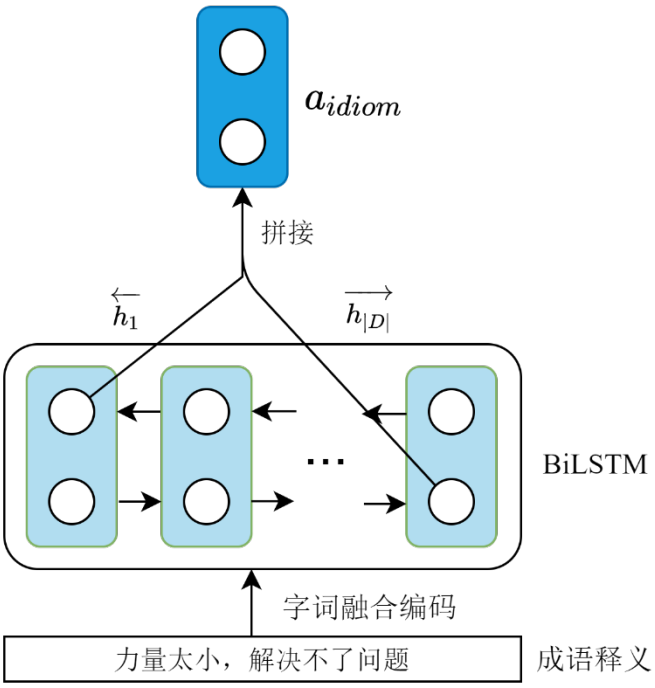


图 5-3 对照组 1 模型结构

Figure 5-3 Architecture of Control Group 1

$$\boldsymbol{a}_{idiom} = \overrightarrow{\boldsymbol{h}_{|D|}} \,||\, \overleftarrow{\boldsymbol{h}_1} \tag{5-7}$$

（2）对 BiLSTM 所有的隐状态向量赋予相同的权重，即对所有的隐状态求和取平均，该对照组是为了验证注意力机制的有效性，同时也是为了验证细粒度信息筛选的必要性，如式 5-8 所示，模型流程图如图 5-4 所示。

$$\boldsymbol{a}_{idiom} = \frac{1}{|D|}\sum_{t=1}^{|D|} \boldsymbol{h}_t \tag{5-8}$$

由三者流程图可以看出，对照组未使用成语的词向量，仅根据其释义信息进行成语的表征，而图 5-1 则使用了词向量作为注意力机制的查询部分的输入，用以与释义中不同词进行交互，动态确定其权重。

5.3 实验验证

5.3.1 实验设置

本节实验验证成语释义增强模型的效果。本章主要针对成语的编码，为了避免上下文编码的影响，以及便于比较对照组和实验组的性能提升效果，实验中上下文编码使用两种编码：

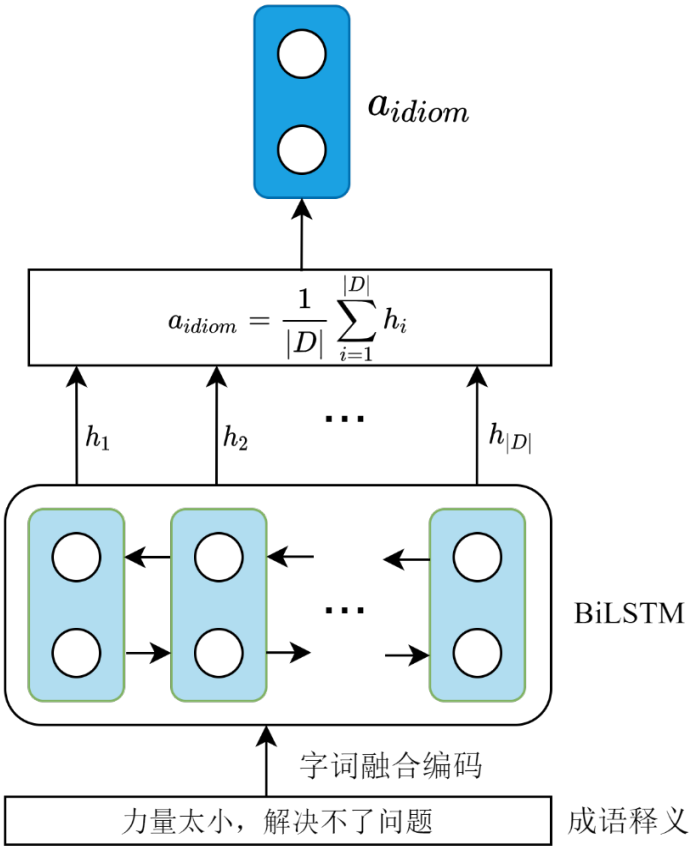


图 5-4 对照组 2 模型结构

Figure 5-4 Architecture of Control Group 2

（1）词向量编码：该组编码与基线模型相同，便于比较成语释义增强模型相对于基线模型的提升效果；

（2）字词融合编码：该组实验是为了验证本文针对现代汉语和成语分别提出的两种编码方式的耦合性能，探究最佳的性能提升效果。

因此实验共两组上下文编码，三组成语编码，共六组实验。

5.3.2 实验结果

六组实验结果如表 5-2 所示。从表中可以得出如下结论：

（1）从性能提升方面分析：“字词融合+释义增强”的性能表现最佳，相较于基线模型性能最高提升了 9.5%（Test 测试集在 AR 模型的表现由 69.6 提升到了 76.2），而其他模型与测试集的性能也提升 4.5%至 9.5%，因此本章提出的多粒度成

语表征有效的提升了成语表征的质量，且与上下文表征耦合性也较好，从而大大提升了模型的整体性能表现。

表 5-2 成语释义增强模型表现

Table 5-2 Performance of Definition-augmented Embedding for Idioms

	Dev	Test	Ran	Sim	Out
基线模型：词+词					
BiLSTM	68.8	68.7	77.9	64.3	59.5
AR	69.8	69.6	79.2	65.7	60.9
SAR	68.9	68.7	78.2	64.4	60.1
词+释义增强					
BiLSTM	73.1	73.2	79.6	66.5	62.9
AR	74.7	74.7	81.5	68.6	64.9
SAR	73.2	73.5	80.2	67.1	63.3
字词融合+释义增强					
BiLSTM	74.5	74.5	81.4	67.8	63.5
AR	**76.1**	**76.2**	**83.8**	**69.9**	**64.9**
SAR	75.2	75.1	82.6	68.1	63.7
词+对照组 1					
BiLSTM	68.9	68.8	77.1	62.1	58.1
AR	69.5	69.6	78.7	63.4	58.9
SAR	69.0	68.9	77.5	62.2	57.9
字词融合+对照组 1					
BiLSTM	67.8	67.6	76.7	62.2	56.2
AR	68.2	68.2	77.5	63.7	56.8
SAR	68.1	68.0	76.3	63.1	55.6
词+对照组 2					
BiLSTM	65.2	65.3	73.5	60.3	55.1
AR	66.3	66.4	74.3	61.1	57.2
SAR	64.7	63.3	73.4	59.3	55.4
字词融合+对照组 2					
BiLSTM	64.1	64.3	71.1	58.3	53.2
AR	65.8	65.7	72.9	59.2	54.7
SAR	64.6	63.9	71.4	59.1	54.1

（2）从上下文表征与成语表征的耦合性分析：从“词+释义增强”和“字词融合+释义增强”两组实验结果中可以看出，对于释义增强成语表征，将上下文表征从词表征变为字词融合后性能有所提升，即上下文字词融合表征与成语的释义增强表征具有较好的耦合性。而从“词+对照组 1”、“字词融合+对照组 1”和“词+对照组 2”、“字词融合+对照组 2”两组对比实验的结果中可以看出，上下文表征

从词表征变为字词融合时，两个对照组的性能表现均有所下降，即对照组的耦合性较差。

（3）从对照组的性能分析：可以看出两个对照组的总体性能均比基线模型差，尤其是对照组 2 性能出现了较大的下降。这是因为两个对照组仅使用了细粒度的信息，且不对这些细粒度信息进行筛选，而加入了注意力机制进行筛选的释义增强则实现了对细粒度信息的有效筛选过滤，即通过成语词向量对不同部分的细粒度信息赋予不同的权重，词向量的语义粒度更“粗”，即包含了更多的整体语义信息，这也侧面印证了成语的意义整体性。对照组 1 性能差的另一个原因是正确答案的语义信息一般只依赖于空所在的位置的句子范围内的上下文，即语义相关的依赖距离较短，而两个方向的 LSTM 的最后一个位置距离空的位置一般较远，尽管 LSTM 的门机制能够实现信息传递过程中的更新，但在较长距离的信息更新后，近距离的语义信息会有所损失，且会有远距离不相关的语义信息加入造成干扰；对照组 2 性能较差的原因与第 4 章三种融合方法的元素加原理相似，此处不再赘述。

5.3.2 权重分析

为了更直观地分析注意力机制对成语释义不同部分的权重分配，在训练好的模型中进行测试时，在式 5-5 步骤取出权重值s_t，以成语“不知深浅”、“满面春风”和“多才多艺”为例，绘制出成语释义不同部分权重分布的热力图，如图 5-5 所示。

热力图右侧为权重数值s_t与颜色的映射关系，权重从 0.0 到 1.0 增大时热力图中对应的方块颜色逐渐加深，颜色越深表示权重越大，即对成语语义的贡献程度越大。从图中可以看出，与成语的语义有关的部分颜色较深，成语“不知深浅”中颜色较深的词语为[“深浅”，“说话”，“做事”，“没有”，“分寸”]，“满面春风”的为[“喜悦”，“笑容”]，“多才多艺”的为[“多方面”，“才能”，“技艺”]；而对于相关性较小的部分，则会赋予较小的权重，如“形容”、“具有”、“或”和标点符号，从而过滤掉这些无用信息。

热力图右侧为权重数值s_t与颜色的映射关系，权重从 0.0 到 1.0 增大时热力图中对应的方块颜色逐渐加深，颜色越深表示权重越大，即对成语语义的贡献程度越大。从图中可以看出，与成语的语义有关的部分颜色较深，成语“不知深浅”中颜色较深的词语为[“深浅”，“说话”，“做事”，“没有”，“分寸”]，“满面春风”的为[“喜悦”，“笑容”]，“多才多艺”的为[“多方面”，“才能”，“技艺”]；而对于相关性较小的部分，则会赋予较小的权重，如“形容”、“具有”、“或”和标点符号，从而过滤掉这些无用信息。

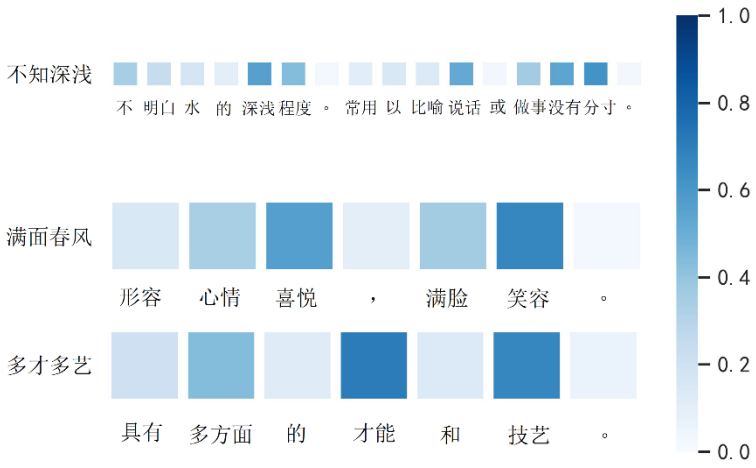


图 5-5 成语释义权重分布

Figure 5-5 The weight distribution of idiom definition

成语表征考虑进释义信息之后，候选答案与上下文进行交互时，使得阅读理解模型能够更加精准的捕获上下文中关联性较强的词语，表 5-3 为成语“不知深浅”的一个案例，上下文中的“艺术品收藏行当‘水很深’”与释义中的“程度深浅”关联性很强，在模型训练的梯度反向传播过程中，阅读理解模型中的注意力机制就能动态地从上下文中选择出与正确答案关联性较强的部分。

表 5-3 权重分析数据案例

Table 5-3 Example of Weight Analysis

列名	数据
content	"据了解，艺术品收藏行当“水很深”，市场价格也是风云莫测，变化颇多，没有几年的积累和摸爬滚打很难摸清其中的门道。经常会有一些小藏家就是在#idiom#、云遮雾障的情况下，被不诚信的卖家忽悠，把假当真，交了高昂的学费。“正是基于这样的一个市场现状，我想搭建出一个以诚信为本的收藏交易的平台。”据任培成介绍，于 2006 年上线试水的博宝网，成立后短短 8 个月时间，日访问量就已达 45 万人次。"
candidates	[["望眼欲穿", "不知深浅", "骨瘦如柴", "不识好歹", "殚思极虑", "不知好歹", "不识时务"]]
groundTruth	["不知深浅"]
realCount	1

5.4 本章小结

本章介绍了针对成语提出的基于释义增强的多粒度成语表征模型。该模型通过基于成语词向量作为查询向量注意力机制，实现了对成语释义数据中细粒度数据的有效筛选，从而有效的利用了成语的多粒度的信息。同时，为了验证两种编码的耦合性以及成语释义信息筛选对的必要性，本章又设置了两个对照组，分别用于模拟仅考虑细粒度信息和不进行细粒度信息筛选的情况。实验结果表明，本章提出的成语释义增强模型与字词融合的上下文表征模型耦合性较好，极大的提升了模型的性能表现，性能最高提升了 9.5%。

6 具体案例与应用分析

本章主要介绍具体案例分析和应用分析。首先，根据具体案例对本文提出的模型的效果进行分析，并通过具体的数学度量进行量化分析；然后，通过收集高考语文试题中成语相关的数据，验证本文提出的模型的实际应用表现。

6.1 度量标准

本节介绍数学领域对向量差异的相关度量标准，包括衡量向量之间直线距离的欧式距离和衡量向量角度大小的余弦相似度。

6.1.1 欧氏距离

欧氏距离即欧几里得度量（Euclidean metric），是一个通常采用的距离定义，指在 m 维空间中两个点之间的真实（直线）距离，或者向量的自然长度（即该点到原点的距离）。对于向量 A、B，其欧氏距离如式 6-2 所示。

$$\mathrm{d}(\boldsymbol{A},\boldsymbol{B}) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \tag{6-2}$$

6.1.2 余弦相似度

在机器学习问题中，在分析两个向量之间的相似性时，常用余弦相似度来表示。余弦相似度通过测量两个向量的夹角的余弦值来度量它们之间的相似度，取值范围是[-1,1]。

具体来说，余弦相似度是使用两个向量之间夹角的余弦值来确定两个向量是否大致指向相同的方向：

（1）当两个向量有相同的指向时，余弦相似度的值为 1；

（2）两个向量夹角为 90 时，余弦相似度的值为 0；

（3）两个向量指向完全相反的方向时，余弦相似度的值为-1。

余弦相似度的值与向量的长度无关，仅仅与向量的指向方向相关。换句话说，余弦相似度关注的是向量之间的角度关系，并不关心它们的绝对大小。对于向量 A、B，其余弦相似度如式 6-1 所示，其中，A_i和B_i分别代表向量 A 和 B 的各维度分量，n 为向量的维度。

$$\cos(\boldsymbol{A},\boldsymbol{B}) = \frac{\boldsymbol{A}\cdot\boldsymbol{B}}{||\boldsymbol{A}||||\boldsymbol{B}||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{6-1}$$

在度量一对文本相似度时，当文本长度差距很大，但内容相近时，如果使用词频或词向量作为特征，它们在特征空间中的欧氏距离可能会很大；但是它们之间的夹角可能会很小，因而相似度很高。此外，余弦相似度在特征维度很高时依然保持着取值在[−1,1]之间的特性，即余弦相似度更注重维度之间的差异，而不注重数值上的差异。总体来说，欧氏距离体现数值上的绝对差异，而余弦距离体现方向上的相对差异。

6.2 错题分析

本节选取正例和负例进行分析，基于正例验证本文提出的理论的效果，并通过负例分析成语表征中仍存在的挑战。

6.2.1 正例分析

表 6-1 为正例数据示例，其中正确答案“肃然起敬”的相似项为“心悦诚服”、“顶礼膜拜”和“叹为观止”，其余的为随机项。

表 6-1 正例数据示例

Table 6-1 Positive Example

列名	数据
content	"我扭过头，看见萧珊的杯子空了。我知道她是吞下去了的，一时#idiom#。本来我对她不做任何要求的。我知道北方的风俗，对女性多有约束——就是现在也是这样。譬如我有石家庄两口子老同学——两口子都是我的大学同班同学，几年前我出差去那里，吃饭的时候，作为妻子的女同学居然就是不上桌，端个碗在旁边走来走去的答腔，把我急得要死。我说你们在学校的时候不是这个样子啊！男同学哈哈大笑，说你不要为难她了，她要坐上来更难受的。"
candidates	[["心悦诚服", "男女老幼", "肃然起敬", "秋高气爽", "顶礼膜拜", "叹为观止", "张三李四"]]
groundTruth	["肃然起敬"]
realCount	1

3.3 节的基线模型中使用了成语向量对成语进行表征，同时在第 5 章对成语表征进行基于释义信息的增强。为了确认释义增强是否能更加有效区分相似成语，对

与两种表征的结果，分别计算正确答案与各候选答案之间的欧式距离和余弦相似度，结果如表 6-2 所示，其中“心悦诚服”、“顶礼膜拜”和“叹为观止”为与相似项。

表 6-2 正例侯选答案量化分析

Table 6-2 Quantitative analysis of Candidate Answers of Positive Example

	心悦诚服	男女老幼	秋高气爽	**顶礼膜拜**	**叹为观止**	张三李四
成语词表征						
欧式距离	**2.85**	3.90	4.37	**2.92**	**2.94**	3.85
余弦相似度	**0.69**	0.41	0.32	**0.66**	**0.66**	0.45
成语释义增强						
欧式距离	**2.91**	3.92	4.35	**3.01**	**2.97**	3.91
余弦相似度	**0.61**	0.42	0.34	**0.62**	**0.61**	0.44

从表 6-2 中可以看出，在词表征中，相似成语之间的余弦相似度较随机项的大，而余弦相似度大意味着其夹角小，另外，可以看出相似成语的欧氏距离较随机项的小。这说明相似成语不但夹角小，而且其距离也相近，因此相似成语是更强的干扰项。

经过释义增强后，三个相似成语与正确成语之间的余弦相似度均有所减小，同时欧式距离也有小幅度的增加，即向量之间的夹角变大了，距离变远了。因此，本文提出的成语释义增强模型能够有效地对相似成语进行判断。

6.2.2 负例分析

表 6-3 为负例数据示例，其中正确答案“一落千丈”的相似项为“一败涂地”、“过街老鼠”和“雪上加霜”，模型的预测结果为“一败涂地”。

使用相同的计算方法算出负例的欧氏距离和余弦相似度，结果如表 6-4 所示。从表中可以看出，经过成语释义增强后，三个相似项的欧式距离只有小幅度提升，余弦相似度也只有小幅度的减小。

为了更直观地感受相似项的迷惑性，我们通过对正确答案和预测答案的释义进行定性分析，如表 6-5 所示。如表中粗体部分中可以看出，相似项“一败涂地”释义中的“低落”、“急剧下降”，“过街老鼠”中的“痛恨”、“坏人坏事”，“雪上加霜”中的“遭受灾难”、“损害”、“严重”，与正确成语“一落千丈”释义中的“失败”、“不可收拾”都表达了较强的负面境况，而这些表达负面境况的词向量之间的欧式距离和夹角较小，如“失败”和“低落”的欧式距离为 2.79，余弦相似度为 0.65。

而通过注意力机制赋予这些词汇较大的权重后，相似项和正确答案的表征中都蕴含了较强的表达负面境况的语义信息，这些细粒度的语义相似度的分辨问题仍然是个具有挑战性的任务。

表 6-3 负例数据示例

Table 6-3 Negative Example

列名	数据
content	"此外，据记者了解，国安象征核心地位的 10 号仍然空缺。这些年来国安的“10 号”球衣有些克主，比如去年的 10 号堤亚哥，上个赛季选择 10 号球衣之后，#idiom#。在客战日本的亚冠比赛之后，记者曾经问堤亚哥，是否感受到了 10 号的压力，因为国安队的 10 号有些克主。堤亚哥的回答是，他本不想选择 10 号，但小马丁拿走了 20 号，他只能穿上 10 号。无论堤亚哥是怎样做出的选择，但堤亚哥背了一年的 10 号后，就走人了，而且可能再也不会回到中超的赛场了。在堤亚哥之前，商毅也曾背过 10 号球衣，可惜那个赛季他一直受到伤病困扰。"
candidates	[["丧尽天良", "一败涂地", "卧薪尝胆", "熙来攘往", "过街老鼠", "一落千丈", "雪上加霜"]]
groundTruth	["一落千丈"]
realCount	1

表 6-4 负例侯选答案量化分析

Table 6-4 Quantitative analysis of Candidate Answers of Negative Example

	丧尽天良	**一败涂地**	卧薪尝胆	熙来攘往	**过街老鼠**	**雪上加霜**
成语词表征						
欧式距离	4.02	**2.77**	3.92	3.88	**2.95**	**2.87**
余弦相似度	0.40	**0.68**	0.34	0.35	**0.64**	**0.70**
成语释义增强						
欧式距离	4.23	**2.81**	4.37	3.97	**2.99**	**3.05**
余弦相似度	0.41	**0.67**	0.33	0.32	**0.63**	**0.66**

表 6-5 负例成语释义分析

Table 6-5 Definition analysis of Candidate Answers of Negative Example

成语	释义
一落千丈	形容彻底**失败**，**不可收拾**。
一败涂地	原指琴声骤然**低落**。后常用以形容景况**急剧下降**。
过街老鼠	比喻人人**痛恨**的**坏人坏事**。
雪上加霜	意思是在雪上还加上了一层霜，在一定天气条件下可以发生，常用来比喻接连**遭受灾难**，**损害**愈加**严重**。

6.3 应用分析

本节通过高考数据集验证本文提出的模型的实际表现，以分析其实际应用，具体包括数据集的收集、任务设置和实验验证。

6.3.1 高考成语试题数据

高考语文试卷中对成语知识的考查是一类稳定的题目，主要考察学生对对成语识记、辨析和运用[52]。高考语文卷中主要采用选择题的方式对成语知识进行考察，选择题的类型主要有两种类型：完形填空和判断正误，如表 6-6 所示。

表 6-6 中选项的序号标记为粗体的为正确选项。从表中可以看出，2020 年全国卷Ⅰ的试题类型为完形填空，其每个位置的候选项均包含了相似成语，难度较大；2019 年全国大纲卷则为判断正误的类型，主要考察成语的语义理解。第一种完形填空的类型与本文实验数据类型一致，因此本文主要收集第一种类型的高考成语试题作为测试数据。

经过分析，高考成语试题数据与本文所用数据有较大的区别。

首先是候选项的设置。成语数据集候选项个数为 7，包括正确答案、相似项和随机项，而高考试题中候选项个数为 4，只有正确答案和相似项，无随机项，且同一个空的候选项有重复成语的现象，表 6-6 中的 2020 年全国卷Ⅰ的试题实际上每个空只有两个选项。

为了使高考试题数据与成语数据格式一致，以适应模型结构，需要对试题数据的候选项进行相应的处理。原始高考试题数据每个空的候选项包括一个正确成语和若干个相似项，需要将每个空的候选项个数扩充至 7。设候选项个数为 num_can，规则如下：

（1）从成语列表中随机抽取 7 - num_can 个不重复的成语，若其中包含原有候选项中已有成语，则重新抽取；

（2）按照 3-2 小节中对同义成语的判断方法，确保选取的候选成语不是正确答案的同义词，若是则重新抽取。

其次，文本长度也有较大的区别。高考试题的文本长度较长，多为 200 到 400 字，而本文所使用的成语数据集的文本长度约为 100，相差较大。通过观察，可以看出高考成语题中空与空是相互独立的，其答案仅取决于其所在句子的上下文。因此，按照成语数据集构建时的思路，根据成语所在的句子，对成语试题进行拆分。表 6-6 试题经过拆分后得到三条数据，即表 6-6 中粗体、斜体和普通文本的部分，三条数据的“realCount”分别为 1、2、1。

表 6-6 高考语文成语试题示例

Table 6-6 Examples of Chinese Idioms Test Questions for College Entrance Examination

2020 年全国卷I
阅读下面的文字，完成下面小题。 **在中国各种艺术形式中，篆刻是一个_____的门类。篆刻是从实用印章的应用中发展而来的，中国的印章最初用在制陶工艺方面，上面镌刻的是图案、花纹或族徽，到春秋战国时期，刻有官职名或人名的文字印章得到普遍使用，唐宋以后，由于文人士大夫参与到印章的创作中，这门从前主要由工匠承揽的技艺，增加了人文意味，印章不再局限于用来昭示身份与权力，而是通过镌刻人名字号，斋馆名称、成语警句等来表达情趣志向，印章也就超越实用功能，成为文人表达自己审美追求的独特方式。***中国印章艺术由此实现了一次完美的升华——演变为中国文化特有的篆刻艺术。明清时期，众多_____的艺术家在篆刻上融入了对汉字形体的研究和理解，再加上他们对印面布局的精心设计,对各种刀法的熟练掌握,篆刻艺术迅速走向成熟并孕育出_____的流派风格。***篆刻艺术的发展及成就，使印章成为与中国画、中国书法紧密结合的艺术形式，同时也是中国画和书法作品中_____的组成部分。 17.一次填入文中横线上的词语，全部恰当的一项是（ ） A.**别具匠心** *才思敏捷 异彩纷呈* 弥足珍贵 B.**别具匠心** *才华横溢 奇光异彩* 不可或缺 **C**.**十分独特** *才华横溢 异彩纷呈* 不可或缺 D.**十分独特** *才思敏捷 奇光异彩* 弥足珍贵
2019 年全国大纲卷
2.以下各句中，划线的成语使用恰当的一项为哪一项 （） A、该产品的试用效果非常好，相信它大量投产后将不负众望，公司一定会凭借产品的优异品质在激烈的市场竞争中取得骄人业绩。 B、某市两家报社相继推出的立体报纸受到广大市民的热烈追捧，更多的立体报纸呼之欲出，可能会成为当地报业的一种发展趋势。 C、中国古典家具曾经非常受消费者青睐，后来很长一段时间市场上却没有了踪影，而在全球崇古风气盛行的今天，它又渐入佳境了。 **D**、这位专家的回答让我有一种醍醐灌顶的感觉，实在没想到这个困扰我两年的问题他却理解得那么轻松。

在数据量上，由于高考真题数据量有限且并不是所有的高考试题都符合要求，因此除了高考真题外，本文还选取了各地的模拟卷、统考试卷、冲刺卷等试卷质量与高考真题较为接近的试题，最后得到了 1000 条测试数据。

在问题的难度上，由于高考成语试题的以选择题形式出现，而且选项只有 4 个，考生可以根据相关性得出正确答案，并不需要作对全部的空。例如，对于表 6-6 的 2020 年全国卷Ⅰ，如果某考生确定了第 1、4 个空的答案，于是便可以选出正确答案 C，而确定第 2、3 个空的答案，这对于考生来说相当于降低了问题的难度。而本文在清洗高考成语试题数据的时候，对试题进行拆分，然后将每个空的候选项单

独扩充至 7 以适应本文提出的模型，这也去除了原题中空之间的相关性了。因此，本文在高考成语试题测试集上的任务难度要大于实际的任务难度。

这 1000 条数据的统计信息如表 6-7 所示，与成语数据集（表 3-5）相比，高考成语试题数据集的平均长度（114）介于领域内数据（99）与领域外数据（127）之间；覆盖成语个数（3176）较成语数据集少；每个文章的平均问题个数、单空比例和多空比例都大致相同。

表 6-7 高考成语试题数据集统计分析

Table 6-7 Statistics of the Dataset of Chinese Idioms Test Questions for College Entrance Examination

统计量	数值
文章个数	1000
平均长度/文章	114
覆盖成语个数	3176
问题总个数	1245
问题平均个数/文章	1.25
单空比例	80.3%
多空比例	19.7%

6.3.2 实验验证

本节介绍本章提出的模型在高考成语试题数据集中的表现，模型使用基线模型和 5.2 节中“字词融合+释义增强”模型，模型训练好以后，将高考成语试题按照 8:2 的比例拆分后，使用 800 条数据对模型进行微调，200 条数据进行测试，进行十次测试取其平均值。结果如表 6-8 所示，这里同时将其他测试集结果列出作为对比。

表 6-8 高考成语试题测试集表现

Table 6-8 Performance of Chinese Idioms Test Questions for College Entrance Examination

	Dev	Test	Ran	Sim	Out	高考
字词融合+释义增强						
BiLSTM	74.5	74.5	81.4	67.8	63.5	**74.1**
AR	76.1	76.2	83.8	69.9	64.9	**75.9**
SAR	75.2	75.1	82.6	68.1	63.7	**74.5**

从表中可以看出，高考成语试题测试集的结果与 Dev 和 Test 测试集结果较为接近，因为高考测试集候选项中虽然相似项个数更少，但是其覆盖的领域更加多样，

因此其综合难度与 Dev 和 Test 较为接近。

高考语文满分 150 分，每年的平均分约为 100 分，如 2019、2020 年湖南省高考语文平均分分别为 100.99、100.63 分[53]，以该平均分可以大致得出成语试题的正确率约为 66.7%，高考语文试卷中不同题目难度不同，因此将该正确率只作为大致参考，真实正确率与此数值会有出入。从本文实验结果中可以看出，在高考成语试题测试集上，本文提出的成语多粒度表征模型的准确率最高达到了 75.9%，相比于 66.7%具有极大的优势，这说明从学生的平均水平来说，本文提出的模型的表现已经好于考生的平均水平，但是仍然有很大的进步空间。

6.4 本章小结

本章通过分析具体的正例和负例对本文提出的模型的性能表现进行具体的量化分析，然后通过收集高考语文试题数据集作为额外的测试集，验证模型的实际应用效果。从这两方面的分析可以得出结论：本文提出的成语多粒度表征模型可以有效地提升成语表征的质量，且在实际应用中其性能表现已经高于考生的平均水平。

7 总结展望

7.1 总结

本文提出了基于释义增强的中文成语多粒度表征模型，并基于完形填空式的中文阅读理解任务验证表征效果，最后将其应用于人工智能高考语文成语试题的解题中，获得了较好的效果。

本文提出了两种表征模型：针对上下文表征的字词融合模型和针对成语表征的释义增强模型。具体来说，字词融合表征中为了解决字向量和词向量数量不一致的问题，本文提了两种对齐模型，即字向词对齐和词向字对齐，然后提出三种融合方式，即元素加、元素乘和拼接；基于释义增强的成语表征中，由于成语的非语义合成性和意义整体性，字信息无法对成语词信息进行补充，为了解决这个问题，本文引入了成语白话文释义，通过注意力机制神经网络对释义信息进行有效的筛选，实现了成语多粒度信息有效结合。最终实验结果表明，本文提出的“字词融合+释义增强”的表征模型在各个测试集中的性能提升了 4.5%至 9.5%。

本文通过对具体案例进行分析，实验验证了模型的实际应用效果。在具体案例的分析中，本文采用欧氏距离和余弦相似度，测量了候选项中各个成语与正确成语之间的距离，通过对比基线模型和本文模型的距离度量，发现本文提出的模型中相似成语的表征之间的欧式距离更大，且余弦相似度更小，即夹角更大，因此可以得出结论：本文提出的成语表征模型对相似成语的辨别能力相较于基线模型更强，从而使模型的性能得到提升。

为了验证本文模型的实际应用效果，本文最后收集了历年高考语文试卷中成语相关的试题，作为测试集，验证模型的实际应用能力。实验结果显示，本文提出的模型在高考测试集上的准确率为 75.9%，极大的高于考生平均水准 66.7%。

7.2 展望

本文工作主要针对成语的多粒度表征，通过引入成语的释义信息对其进行信息补充为成语表征提供了新的思路。本文主要针对阅读理解模型中的编码层的部分，未涉及交互层和输出层。而随着预训练模型的兴起，自然语言处理各种任务有了极大效果提升。未来工作中可以考虑引入预训练模型，以增强模型的阅读理解能力，同时也可以对成语释义进行表征。另一方面，成语的另一个应用，即成语的恰当使用也是一个重要的应用方向，未来计划尝试针对此类任务进行相关实验。

参考文献

[1] 秦赞. 中文分词算法的研究与实现[D].吉林大学,2016.

[2] 杨晓宏,周效章.我国在线教育现状考察与发展趋向研究——基于网易公开课等 16 个在线教育平台的分析[J].电化教育研究,2017,38(08):63-69+77.

[3] 徐震. 对外汉语成语教学的文化导入研究[D].浙江大学,2013.[4] 陈健鹏,马建辉,王怡君.基于多轮交互的人机对话系统综述[J].南京信息工程大学学报(自然科学版),2019,11(03):256-268.

[4] 杜锦绣,蔡静. 网络舆情监测的数据采集与文本分类技术分析[J].无线互联科技, 2019, 16(15):123-124.

[5] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[6] Le Q, Mikolov T. Distributed representations of sentences and documents[C]//International conference on machine learning. pages 1188-1196.

[7] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation [C] // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pages 1532-1543.

[8] Wang S. Chinese Multiword Expressions: Theoretical and Practical Perspectives. Springer Singapore, 2019.

[9] Joulin A, Grave E, Bojanowski P, et al. Bag of tricks for efficient text classification[J]. arXiv preprint arXiv:1607.01759, 2016.

[10] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.

[11] Lai S, Xu L, Liu K, et al. Recurrent convolutional neural networks for text classification[C]//Twenty-ninth AAAI conference on artificial intelligence. 2015.

[12] Li S, Zhao Z, Hu R, et al. Analogical Reasoning on Chinese Morphological and Semantic Relations, ACL 2018.

[13] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[14] Radford A, Narasimhan K. Improving Language Understanding by Generative Pre-Training. 2018.

[15] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019.

[16] Dai Z, Yang Z, Yang Y, et al. Transformer-xl: Attentive language models beyond a fixed-length context[J]. arXiv preprint arXiv:1901.02860, 2019.

[17] Sun Y, Wang S, Li Y, et al. ERNIE: Enhanced Representation through Knowledge Integration[J]. arXiv preprint arXiv:1904.09223, 2019.

[18] Zhang Z, Han X, Liu Z, et al. ERNIE: Enhanced Language Representation with Informative Entities[J]. arXiv preprint arXiv:1905.07129, 2019.

[19] Song K, Tan X, Qin T, et al. Mass: Masked sequence to sequence pre-training for language generation[J]. arXiv preprint arXiv:1905.02450, 2019.

[20] Yang Z, Dai Z, Yang Y, et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding[J]. arXiv preprint arXiv:1906.08237, 2019.

[21] Wenzhen J, Hong Z, Guocai Y. An Efficient Character-Level and Word-Level Feature Fusion Method for Chinese Text Classification[C]//Journal of Physics: Conference Series. IOP Publishing, 2019, 1229(1): 012057.

[22] Li J, Wan X, Qin S. Word-Level and Character-Level Mixed Features for Chinese Short Text Classification[C]//2018 IEEE 4th International Conference on Computer and Communications (ICCC). IEEE, 2018: 2344-2348.

[23] Bojanowski P, Grave E, Joulin A, et al. Enriching word vectors with subword information[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146.

[24] Zhang Z, Huang Y, Zhao H. Subword-augmented embedding for cloze reading comprehension[J]. arXiv preprint arXiv:1806.09103, 2018.

[25] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[26] Dou Z, Zhang Z. Hierarchical Attention: What Really Counts in Various NLP Tasks[J]. arXiv preprint arXiv:1808.03728, 2018.

[27] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[28] Cui Y, Che W, Liu T, et al. 2019. Pre-training with whole word masking for Chinese bert. arXiv preprint arXiv:1906.08101.

[29] He W, Liu K, Liu J, et al. 2018. Dureader: A Chinese machine reading comprehension dataset from real-world applications. In Proceedings of the Workshop on Machine Reading for Question Answering@ACL 2018, pages 37–46, Melbourne, Australia.

[30] Sun K, Yu D, Yu D, Cardie C. Investigating prior knowledge for challenging Chinese machine reading comprehension. Transactions of the Association for Computational Linguistics. 2020;8:141-55.

[31] Cui Y, Liu T, Chen Z, Ma W, Wang S, Hu G. Dataset for the first evaluation on Chinese machine reading comprehension. arXiv preprint arXiv:1709.08299. 2017 Sep 25.

[32] Cui Y, Liu T, Chen Z, Wang S, Hu G. Consensus attention-based neural networks for Chinese reading comprehension. arXiv preprint arXiv:1607.02250. 2016 Jul 8.

[33] Wang L, Yu S. Construction of Chinese Idiom Knowledge-base and Its Applications[C]//Proceedings of the 2010 Workshop on Multiword Expressions: From Theory to Applications. 2010: 11-18.

[34] Wang M, Xiao M, Li C, et al. STAC: Science Toolkit Based on Chinese Idiom Knowledge Graph[C]//Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications. 2019: 57-61.

[35] Zheng C, Huang M, Sun A. ChID: A Large-scale Chinese Idiom Dataset for Cloze Test[J]. arXiv preprint arXiv:1906.01265, 2019.

[36] Jiang Z, Zhang B, Huang L, et al. Chengyu Cloze Test[C]//Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications. 2018: 154-158.

[37] Liu Y, Pang B, Liu B. Neural-based Chinese Idiom Recommendation for Enhancing Elegance in Essay Writing[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pages 5522-5526.

[38] Muzny G, Zettlemoyer L. Automatic idiom identification in Wiktionary. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing 2013 Oct (pp. 1417-1421).

[39] Long S, Wang R, Tao K, Zeng J, Dai XY. Synonym Knowledge Enhanced Reader for Chinese Idiom Reading Comprehension. arXiv preprint arXiv:2011.04499. 2020 Nov 9.

[40] Tan M, Jiang J. A BERT-based Dual Embedding Model for Chinese Idiom Prediction. arXiv preprint arXiv:2011.02378. 2020 Nov 4.

[41] Liu P, Qian K, Qiu X, Huang XJ. Idiom-aware compositional distributed semantics. In Proceedings of the 2017 conference on empirical methods in natural language processing 2017 Sep (pp. 1204-1213).

[42] Shao Y, Sennrich R, Webber B, Fancellu F. Evaluating machine translation performance on Chinese idioms with a blacklist method. arXiv preprint arXiv:1711.07646. 2017 Nov 21.

[43] Pelletier FJ. The principle of semantic compositionality. Topoi. 1994 Mar 1;13(1):11-24.

[44] Shao J. General introduction to modern Chinese. Shanghai, China: Shanghai Educational Publishing House, 2018.

[45] 国学网.《成语大全》. http://www.guoxue.com/chengyu/CYML.htm

[46] Fleiss' Kappa. https://www.real-statistics.com/reliability/interrater-reliability/fleiss-kappa/

[47] Song Y, Shi S, Li J, Zhang H. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) 2018 Jun (pp. 175-180).

[48] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997 Nov 15;9(8):1735-80.

[49] Hermann KM, Kočiský T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P. Teaching machines to read and comprehend. arXiv preprint arXiv:1506.03340. 2015 Jun 10.

[50] Chen D, Bolton J, Manning CD. A thorough examination of the CNN/daily mail reading comprehension task. arXiv preprint arXiv:1606.02858. 2016 Jun 9.

[51] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014 Dec 22.

[52] 葛莉茜. 新课标背景下高考成语辨析题解析及备考策略——以 2018 年高考全国卷为例[J]. 福建教育学院学报,2019,20(02):125-128.

[53] 好老师教育. 2020 高考各科平均分. https://www.sohu.com/a/409551819_120068693 ?trans_=000014_bdss_dkqgad

作者简历

一、作者简历

李想，男，1994 年 3 月生。2014 年 9 月至 2018 年 6 月就读于北京交通大学电子与信息工程学院通信工程专业，取得工学学士学位。2018 年 9 月至 2021 年 6 月就读于北京交通大学电子与信息工程学院通信与信息系统专业，研究方向是自然语言处理，取得工学硕士学位。攻读硕士学位期间，主要从事自然语言处理的工作。

二、发表论文

[1] X. Li, Y. Guo, Y. Sheng and Y. Chen, "Characterizing Social Marketing Behavior of E-commerce Celebrities and Predicting Their Value," IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, Canada, 2020, pp. 1166-1171, doi:10.1109/INFOCOMWKSHPS50562.2020.9162757.

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京交通大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名: 签字日期: 年 月 日

学位论文数据集

表 1.1： 数据集页

关键词*	密级*	中图分类号*	UDC	论文资助
自然语言处理；成语表征；中文机器阅读理解	公开			
学位授予单位名称*		学位授予单位代码*	学位类别*	学位级别*
北京交通大学		10004	工学	硕士
论文题名*		并列题名		论文语种*
中文成语表征学习及其应用				汉语
作者姓名*	李想		学号*	18120100
培养单位名称*		培养单位代码*	培养单位地址	邮编
北京交通大学		10004	北京市海淀区西直门外上园村 3 号	100044
学科专业*		研究方向*	学制*	学位授予年*
通信与信息系统		自然语言处理	3	2021
论文提交日期*	2021.05.6			
导师姓名*	郭宇春		职称*	教授
评阅人	答辩委员会主席*		答辩委员会成员	
电子版论文提交格式 文本（√） 图像（ ） 视频（ ） 音频（ ） 多媒体（ ） 其他（ ） 推荐格式：application/msword；application/pdf				
电子版论文出版（发布）者		电子版论文出版（发布）地		权限声明
论文总页数*	49 页			
共 33 项，其中带*为必填数据，为 22 项。				