基于新闻图式结构的篇章功能语用识别方法

杜梦琦[1]，蒋峰[1]，褚晓敏[1]，李培峰[1,2]
[1]苏州大学计算机科学与技术学院，苏州，中国
[2]苏州大学人工智能研究院，苏州，中国
{20205227068，20194027003}@stu.suda.edu.cn，{xmchu，pfli}@suda.edu.cn

摘要

篇章分析是自然语言处理领域的研究热点和重点，篇章功能语用研究旨在分析篇章单元在篇章中的功能和作用，有助于深入理解篇章的主题和内容。目前篇章分析研究以形式语法为主，而篇章作为一个整体的语义单位，其功能和语义却没有引起足够重视。已有功能语用研究以面向事件抽取任务为主，并未进行通用领域的功能语用研究。鉴于功能语用研究的重要性和研究现状，本文提出了基于新闻图式结构的篇章功能语用识别方法来识别篇章功能语用。该方法在获取段落交互信息的同时又融入了篇章的新闻图式结构信息，并结合段落所在篇章中的位置信息，从而有效地提高了篇章功能语用的识别能力。在汉语宏观篇章树库的实验结果证明，本文提出的方法优于所有基准系统。

关键词： 篇章分析；篇章功能语用；新闻图式结构

Discourse Functional Pragmatics Recognition Based on News Schemata

Mengqi Du[1], Feng Jiang[1], Xiaomin Chu[1], Peifeng Li[1,2]
[1]School of Computer Science and Technology, Soochow University, Suzhou, China
[2]AI Research Institute, Soochow University, Suzhou, China
{20205227068，20194027003}@stu.suda.edu.cn，{xmchu，pfli}@suda.edu.cn

Abstract

Discourse analysis is a hot topic in the field of natural language processing. The purpose of discourse functional pragmatics research is to analyze the function and role of discourse units, which is helpful to deeply understand the theme and content of discourse. At present, discourse analysis mainly focuses on formal grammar, but the function and semantics of discourse as a whole semantic unit have not attracted enough attention. The existing functional pragmatics researches are mainly oriented to event extraction task, but there is no general functional pragmatics research. In view of the importance and status of functional pragmatics research, this paper proposes a Functional Pragmatics Recognition Method Based on News Schemata(FPRNS). FPRNS not only obtains the interaction information of paragraphs, but also incorporates the information of news schemata and the location information of paragraphs, so as to effectively improve the recognition ability of discourse functional pragmatics. The experimental results in the Chinese macro discourse tree-bank show that the proposed method is superior to all baselines.

Keywords: Discourse Analysis , Discourse Functional Pragmatics , News Schemata

第二十一届中国计算语言学大会论文集，第120页-第131页，南昌，中国，2022年10月14日至16日。

1 引言

当代语言学有两大流派，分别是形式主义和功能主义。目前自然语言处理领域的相关研究大都基于形式语法，这是由于形式语法的中心任务是研究语法成分之间的形式关系，运用鲜明的数理符号表示，促进了语言处理的可计算化。而功能语法以功能和语义为导向，将篇章作为一个语言使用单位，比较难以进行抽象化，因此目前篇章分析领域缺乏针对篇章整体的功能语用研究。

功能语用研究旨在分析篇章单元在篇章中所承担的角色和所起的作用，有助于挖掘篇章中具有价值的信息，深入理解篇章的主题和表达的含义，可以应用于自然语言处理中的其他任务，包括问答系统、信息抽取、作文自动评分等。

在功能语法的研究方面，Halliday (1994)创立了系统功能语言学，他明确地指出，系统功能语言学在本质上是“功能的”和“语义的”，而不是“形式的”和“句法的”，系统功能语言学的研究对象是“语篇”，而不是“句子”。Van Dijk根据新闻报道的结构和功能特点，提出了新闻图式理论(Van Dijk, 1988)。该理论将篇章研究和媒体研究有机结合起来，集中论述了新闻报道的语用结构，为篇章功能语用的研究奠定了基础。近些年来，研究者基于该理论标注了一系列语料资源。在英文方面，Yarlott et al. (2018)标注了来自ACE2语料库的50篇文章段落的功能语用，但该语料规模较小；Choubey et al. (2020)根据新闻图式理论中的功能语用类型做出调整，定义了8种功能语用类型，标注了802篇新闻中句子的功能语用，但其定义的功能语用类型只适用于事件抽取任务。在中文方面，Chu (2018)在新闻图式理论的基础上将细颗粒度的功能语用类型进行合并，补充了新闻图式理论中没有但在实际新闻报道中出现的功能语用类型，共定义了18种功能语用类型。在此基础上Jiang et al. (2018)标注了规模为720篇的宏观篇章语料库（MCDTB），为中文篇章功能语用研究奠定了基础。

P1:记者日前从云南省民政厅获悉，根据中国、老挝和联合国难民署三方达成的遣返在华老挝难民协议，云南顺利完成了十二批、二千九百一十七人的遣返任务，遣返人数占原在华老挝难民的百分之七十三以上。
P2:根据国务院的统一部署，自一九七八年以来，云南先后接收安置难民六万四千一百余人，其中老挝难民到一九九一年已达三千九百九十四人。依照中国政府缔结的有关国际公约，云南省对难民实行大量的国际主义和人道主义援助。十五年来，在全省尚有四十多个贫困县接受政府财政补助的情况下，已累计支付二点六亿元专为难民解决生产生活困难。
P3:国际社会公认，难民自愿遣返到原籍国是永久解决难民问题的最佳方案。随着中、老两国关系的不断改善，老挝政府主动表示愿意接收在华老挝难民回国，遣返难民的条件日趋成熟。一九九一年四月和七月，中老两国政府正式签订了《关于遣返在华老挝难民的议定书》和《备忘录》，分别委托中国云南省政府和老挝南塔省政府具体负责组织实施。
P4:在遣返过程中，云南省政府对难民的生活极为重视，并同老挝政府和联合国难民署密切配合，使遣返工作得以顺利进行。
P5:据悉，云南尚余一千零七十五名老挝难民，在难民完全自愿前提下，云南省政府将继续积极稳妥进行遣返。（完）

例1 chtb_0255

本文以中文篇章功能语用为研究对象，在MCDTB的基础上开展篇章功能语用的探索与研究。本文以MCDTB中的一篇文章(chtb_0255)来说明篇章的功能语用结构，文章内容如例1所示。其中，段落P1为全文“导语”，阐述了全文的主要内容“云南顺利完成了在华老挝难民的遣返任务”，P2阐述了近年来云南省对难民实行大量的人道主义援助，介绍了文中事件发生的“背景”，P1和P2两个段落形成文章的“总述”部分。P3、P4分别讲述了老挝政府愿意接受难民回国以及在遣返过程中各方密切配合使得遣返任务顺利进行，是文章的“情景”部分，两个段落组成全文的“故事”，同时P5补充了云南省将继续对剩余老挝难民的遣返工作，是全文内容的“补充”。

例1可用图1所示篇章功能语用结构树表示。其中，叶子节点表示新闻报道的段落P1到P5的

基金项目：国家自然科学基金(61836007,61773276);江苏高校优势学科建设工程资助项目

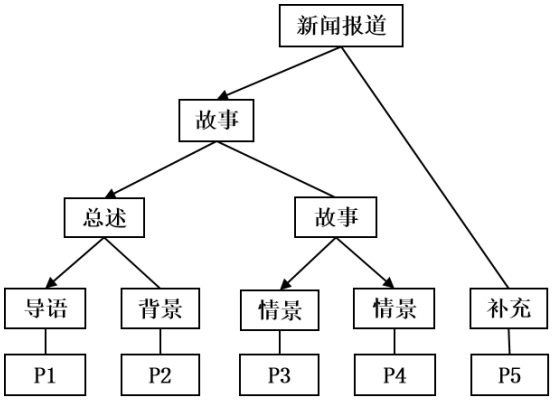

图 1. chtb_0255篇章功能语用结构树

功能语用；非叶子节点表示由其下层篇章单元组成的篇章单元在篇章中的功能语用；根节点表示整个篇章的功能语用。该结构树中的箭头指向重要的篇章单元。

新闻图式理论(Van Dijk, 1988)明确指出单个段落在整篇文章发挥着重要的作用，因此本文将篇章功能语用的识别建模成识别篇章中每个段落的功能语用的任务。结合新闻图式理论，本文分析了新华社的新闻报道，发现新闻中的每一个段落从属于一个更大的功能语用范畴，例如图1中“导语”、“背景”属于“总述”这个功能语用范畴，通过对文章进行范畴划分有助于读者先从整体上理解文章含义，进而更好地把握每一个范畴中段落的功能语用，这样的文章理解思路在Zhang (2014)的语言学研究中也有相应的论述。另一方面，我们研究了新闻报道本身的特点，比较规范的新闻报道文本，第一段功能语用常为“导语”，交代新闻事件的要素，最后一段则常常是对整篇新闻报道的“评论”或“补充”，帮助读者厘清事件的意义或补充事件后续发展。因此，段落在篇章中的位置信息对于识别篇章的功能语用具有重要的作用。

为了深入研究篇章功能语用，本文提出了一个基于新闻图式结构的篇章功能语用识别模型。该模型首先将段落通过XLNet获得段落初步编码，然后通过指针网络获得段落交互信息和篇章范畴划分信息，从而获得篇章组织结构信息。此外，该模型又结合段落在篇章中的位置信息，获得了更加丰富的段落表示，从而有效地提高了篇章功能语用的识别能力。在MCDTB的实验结果表明，本文提出的模型对篇章功能语用的识别有很好的效果。

2 相关工作

目前自然语言处理领域的相关研究大都基于形式语法，以研究语言的结构和形式为主要任务，这是由于形式语法运用鲜明的数理符号表示，而功能语法将篇章作为一个语言使用单位，比较难以进行抽象化和形式化，因此目前篇章分析领域缺乏针对篇章整体的功能语用研究。

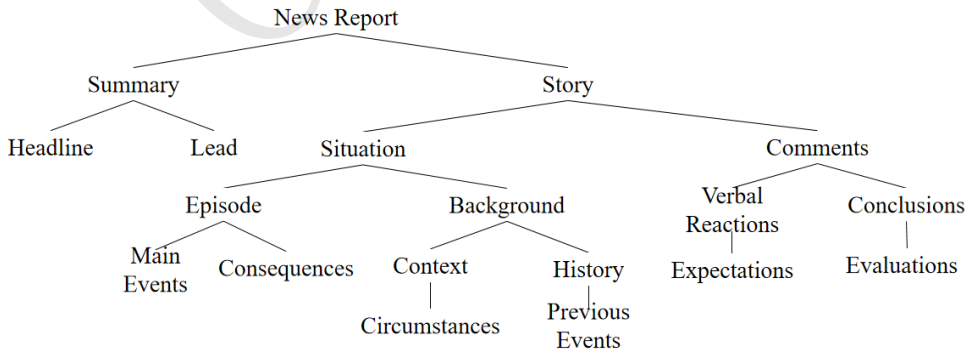

图 2. 假拟新闻图式结构

现有的涉及篇章功能语用的理论包括Van Dijk的新闻图式理论(Van Dijk, 1988)和Pan and Kosicki (1993) 提出的基于框架的方法。新闻图式理论(Van Dijk, 1988)将篇章研究和媒体研究有机结合起来，集中论述了新闻报道的语用结构。图2是一个假拟的新闻图式结构，该结构包含

总述和故事两个最主要的部分。其中，总述（Summary）由标题（Headline）和导语（Lead）两个部分组成，故事（Story）由情景（Situation）与评论（Comments）组成，新闻图式结构通过一个从上而下的层级顺序清晰地展示了新闻篇章的整体组织形式。Pan and Kosicki (1993)提出的基于框架的方法从四个维度分析新闻篇章：句法结构、脚本结构、主题结构和修辞结构，其中句法结构与新闻图式理论最为相似。

针对英文篇章功能语用的研究，Liddy (1991)、Kircz (1991)和Teufel et al. (1999)利用修辞结构和论元类型来定义功能类型并且为科学文章创建了语料库；Mizuta et al. (2006)、Wilbur et al. (2006)、de Waard et al. (2009)和Liakata et al. (2012)利用多种注释模式，对生物领域的功能语用进行了研究。但这些研究针对的是科学论文和生物领域，不适用于其他领域。一些研究者在新闻图式理论基础上，标注了事件领域的语料资源。例如，Yarlott et al. (2018)标注了来自ACE2（Automated Content Extraction Phase2）语料库的50篇文章，并且分别使用SVM、决策树和随机森林等传统机器学习方法做了验证性的实验。而Banisakher et al. (2020)在Yarlott的基础上，利用CRF模型对每个段落的功能语用进行预测，提升了段落功能语用的识别性能。Choubey et al. (2020)将新闻图式理论中的功能语用类型进行相应的调整，定义了8种功能语用类型，标注了802篇文章中句子的功能语用，并提出了一个两层的双向LSTM对文章中每个句子的功能语用进行预测。Choubey and Huang (2021)提出了一个演员-评论家框架来识别句子功能语用，该模型使用了多个评论家，评论家根据已知的子话题结构采取行动，而演员模型的目标是超越评论家，并且引入了一个分层神经网络建模句子、子话题和文档之间的交互。在上述的研究中，Yarlott et al. (2018)标注的语料规模较小，且标注者之间的Kappa值（55%）并不高；Choubey et al. (2020)定义的功能语用类型面向事件，不适用于其他领域。

而中文方面，Song et al. (2020)标注了1230篇作文中句子的功能语用，共定义了7种功能语用类型，并基于序列标注思想，结合句子在作文中的位置信息对句子功能进行预测，但是其标注的功能语用类型是面向学生议论文的，不适用于其他领域；Chu et al. (2018)和Chu (2018)参考宏观架构理论和新闻图式理论提出了一套宏观篇章结构表示体系，在此基础上Jiang et al. (2018)标注了720篇新闻报道，形成宏观汉语篇章树库（Macro Chinese Discourse Treebank, MCDTB）。在MCDTB语料库上，已有的工作都是针对篇章逻辑语义的研究(Jiang et al., 2021; Sun et al., 2020)，对篇章功能语用的研究只进行了初步尝试，研究较少。

3 任务介绍

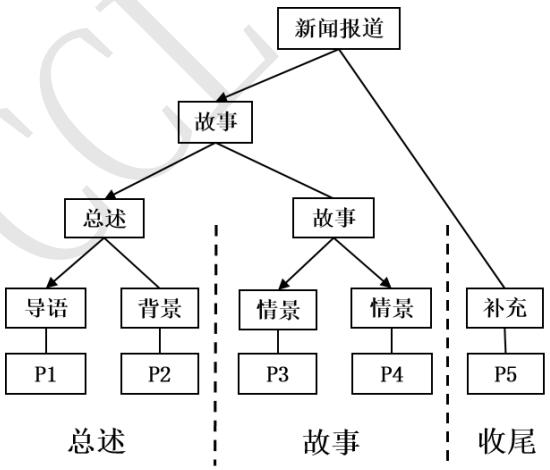

图 3. 新闻报道的范畴划分

一些语言学家在新闻图式理论的基础上又做了进一步阐述。例如，Norman (1995)将新闻图式结构中的“情节”和“评论”更名为“附属”和“收尾”，认为新闻应该由“总述-附属-收尾”三个范畴组成；Bell and Garrett (1998)认为新闻框架应该由“属性-概述-故事”三个范畴组成，有一些新闻还包含“补充”，如背景、评论和追踪报道等，其中“概述”包括“标题”和“导语”，相当于新闻图式结构中的“总述”；Zhang (2014)认为新闻篇章应该由“新闻摘要-新闻故事-新闻评议”三个范畴组成，其中“新闻摘要”对应新闻图式理论的“总述”，“新闻故事”对应“情节”，“新闻评

议”对应“评论”。综合上述理论体系研究和对MCDTB中样本的分析，本文认为“总述-故事-收尾”是新闻报道必备的功能语用结构，因此本文将新闻报道划分为“总述-故事-收尾”三个范畴。

基于上述理论研究和样本分析，本文在MCDTB语料上进行功能语用的范畴划分，形式如图3所示。具体划分方法：（1）将“导语”段以及对“导语”进行解释说明的段落（例如，背景段、补充段等）标注为“总述”；（2）将新闻篇章描述的主体事件标注为“故事”，该部分包括事件主要情节的详细阐述、原因分析、数据支撑等；（3）将对新闻报道事件进行评价、总结或者补充的段落标注为“收尾”。其中，“故事”是新闻的主体部分，是每篇新闻报道的必要部分，而“总述”和“收尾”则可以省略。经过语料处理，48.1%的新闻包含三部分，46.8%包含“总述-故事”两个部分，3.1%包含“故事-收尾”两个部分，还有2.1%仅由“故事”组成。

新闻图式理论(Van Dijk, 1988)指出单个段落在整篇文章发挥着重要的作用，因此本文将段落功能语用作为研究对象，将篇章功能语用识别任务专注于识别叶子节点的功能语用。具体而言，对于给定的一篇文章$T = \{P_1, P_2, ..., P_m\}$，篇章功能语用识别任务就是通过模型识别出段落的功能语用$T_{Fun} = \{Fun_1, Fun_2, ..., Fun_m\}$。本文使用准确率来评判模型对于段落功能语用的识别能力。以chtb_0255为例，正确的段落功能语用为{导语，背景，情景，情景，补充}，若模型的预测结果为{导语，补充，情景，情景，补充}，那么模型预测的准确率为4/5=80%。

4 FPRNS模型

本文提出了一种基于新闻图式结构的篇章功能语用识别模型（Functional Pragmatics Recognition Based on News Schemata，FPRNS），如图4所示。该模型主要由四部分组成：1）文本编码层（Text Encoding Layer）；2）范畴划分层（Category Segmentation Layer）；3）范畴识别层（Category Classifier Layer）；4）信息融合层（Information Fusion Layer）。

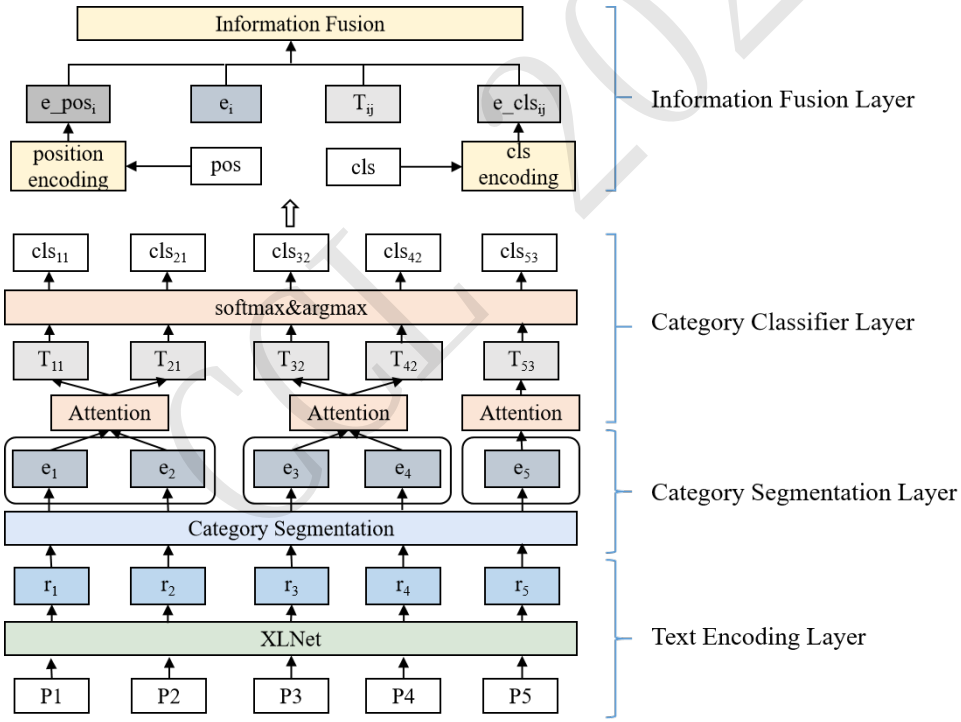

图 4. 篇章功能语用识别模型图

4.1 文本编码层（Text Encoding Layer）

文本编码层使用XLNet作为编码器对段落内容进行编码得到段落向量表示。假设一篇文章含有m个段落，段落序列为$T = \{P_1, P_2, ..., P_m\}$，将每段使用分隔符“SEP”将段落区分开，得到$T' = \{P_1, SEP, P_2, SEP, ..., P_m, SEP, CLS\}$。由于XLNet可处理的数据最大长度为1024，所以需要对段落序列T'进行处理。如果序列长度超过最大长度，则比较序列中每个段落的长度，将最长段落末尾的字符截掉，直至序列长度为最大长度。然后使用XLNet对段落序列T'进

行编码，取每个段落编码最后一个词的词向量作为该段落的语义表示，得到段落语义表示序列$T_r = \{r_1, r_2, ..., r_m\}$。

4.2 范畴划分层（Category Segmentation Layer）

范畴划分层采用指针网络对篇章中功能语用范畴进行划分。具体模型如图5所示。

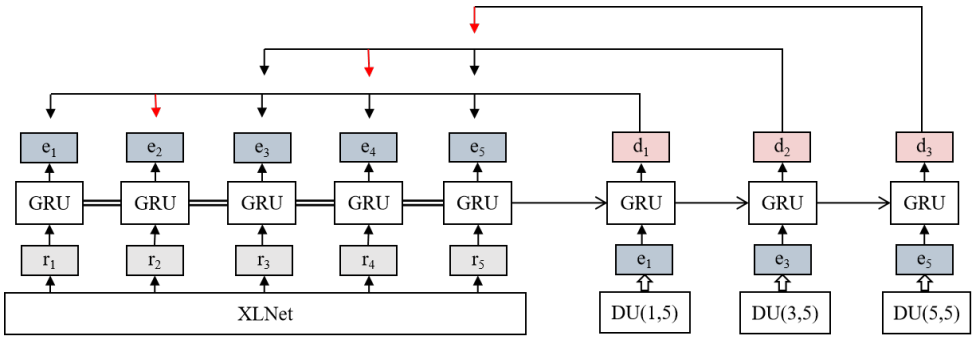

图 5. 范畴划分层模型图

4.2.1 编码层

Chung et al. (2014)的研究表明，GRU(Cho et al., 2014)和LSTM在很多任务上的性能不分伯仲，但是GRU拥有更少的参数，容易收敛，因此在编码层本文使用Bi-GRU进行编码。以chtb_0255为例，本文将段落序列T $= \{P_1, P_2, P_3, P_4, P_5\}$经过文本编码层得到将段落语义表示$T_r$，将$T_r$输入到Bi-GRU中，得到具有交互信息的段落语义表示序列$T_e = \{e_1, e_2, e_3, e_4, e_5\}$，其中$e_i = [e_i^f, e_i^b]$。$e_i^f$,$e_i^b$分别表示正向和反向的输出。

4.2.2 解码层

在解码层采用的也是一个GRU。本文将编码层的输出$T_e = \{e_1, e_2, e_3, e_4, e_5\}$作为解码层的输入。假设在第t步解码时，篇章单元队列为DU(l,5)，解码层会综合当前篇章单元的队头语义表示e_l和t步之前生成的篇章单元语义信息生成当前状态d_t。d_t和当前篇章单元语义信息$T_{e(l,5)} = \{e_l, e_{l+1}, ..., e_5\}$进行交互，通过softmax层得到关于$T_{e(l,5)}$的概率分布。其中$\sigma(.,.)$是融合当前状态表示和篇章单元交互信息，具体为点积运算；α_t为关于$T_{e(l,5)}$的概率分布。如公式（1）所示。

$$\begin{aligned} s_{t,i} &= \sigma(d_t, e_i)\ , i = l, ..., 5 \\ \alpha_t &= softmax(s_{t,i}) \end{aligned} \tag{1}$$

如果通过softmax层后e_i被分配的概率值越大，表明段落P_i和P_{i+1}之间的语义联系越松散，因此这两个段落应该分属于两个功能语用范畴中，从而将篇章划分为两个篇章单元DU(l,i)和DU(i+1,5)。每一步解码，将划分后的两个篇章单元的后者继续放入队列，递归地对篇章单元进行切分，直至队空，解码过程如图6所示。

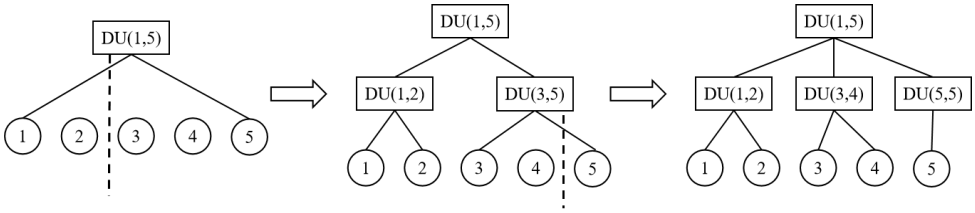

图 6. 解码过程

由章节3可知，每篇新闻报道最多只有三个范畴，但是在实际识别的时候可能会识别出三个以上的范畴。对于这种情况本文比较切分点的概率值，选取前两个概率值较大的切分点作为最终的切分点的位置。

本文采用负对数似然作为损失函数，记为L_1，如公式（2）所示，其中$y_{<t}$表示在第t步解码之前产生的篇章单元，θ为可训练的参数。

$$L_1(\theta) = -\sum_{i=1}^{batch}\sum_{t=1}^{T} logP(y_t|y_{<t}, X) \tag{2}$$

4.3 范畴识别层（Category Classifier Layer）

经过范畴划分层，得到三个范畴，每个范畴包含的段落分别为$\text{Topic}_1 = \{\text{P}_1, \text{P}_2\}$，$\text{Topic}_2 = \{\text{P}_3, \text{P}_4\}$，$\text{Topic}_3 = \{\text{P}_5\}$，范畴语义表示分别为$Topic_{e1} = \{e_1, e_2\}$，$Topic_{e2} = \{e_3, e_4\}$，$Topic_{e3} = \{e_5\}$。将具有上下文交互信息的段落语义表示$e_i$通过注意力机制（即范畴中每个段落语义信息的加权和）来获得范畴的语义表示，如公式（3）所示，将段落所在范畴语义表示序列记为$T_{topic} = \{T_{11}, T_{21}, T_{32}, T_{42}, T_{53}\}$，其中$T_{ij}$表示第i段对应第j个范畴。

$$T_{ij} = Attention(Topic_{ej}) \tag{3}$$

将T_{topic}送入到softmax层获取范畴所属类别的概率分布，从而得到每个范畴具体的类别信息，如公式（4）所示。将范畴类别序列记为$T_{topic_cls} = \{cls_{11}, cls_{21}, cls_{32}, cls_{42}, cls_{53}\}$,其中$cls_{ij}$表示第i段对应第j个范畴类别。

$$cls_{ij} = argmax(softmax(T_{ij})) \tag{4}$$

本文同样采用负对数似然作为范畴识别层的损失函数，记为L_2，如公式（5）所示。其中，θ为可训练的参数。

$$L_2(\theta) = -\frac{1}{N}\sum_{i=1}^{N} log(y\_cls_{ij}), y\_cls_{ij} = softmax(T_{ij}) \tag{5}$$

4.4 信息融合层（Information Fusion Layer）

信息融合层将段落语义表示、范畴语义信息、范畴分类信息以及段落位置信息进行融合，具体模型如图7所示。

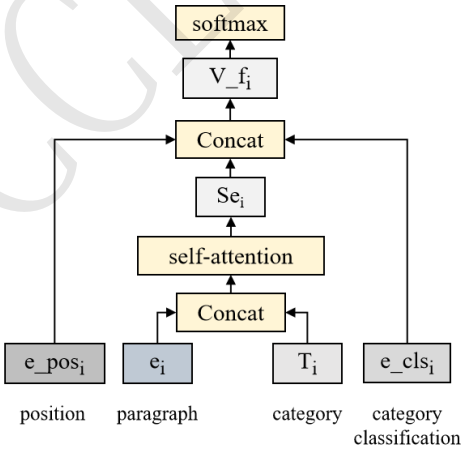

图 7. 信息融合层模型图

将段落交互信息T_e和范畴语义信息T_{topic}通过自注意力机制进行融合，获得更加丰富的段落语义信息表示序列$T_{Se} = \{Se_1, ..., Se_5\}$，其中$Se_i$表示段落交互信息和范畴语义信息融合后的段落表示，如公式（6）所示。注意力机制计算公式如（7）所示。

$$T_{Se} = Attention(concat(T_e, T_{topic})) \tag{6}$$

第二十一届中国计算语言学大会论文集，第120页-第131页，南昌，中国，2022年10月14日至16日。

$$Attention\,(Q,K,V)=softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

将范畴类别信息$T_{topic_cls}=\{cls_{11},cls_{21},cls_{32},cls_{42},cls_{53}\}$当作特征进行编码，得到编码后的范畴类别表示序列$T_{e_cls}=\{e_cls_{11},e_cls_{21},e_cls_{32},e_cls_{42},e_cls_{53}\}$，其中$e_cls_{ij}$表示范畴类别编码后的结果。

本文认为篇章功能语用识别与段落所在篇章中的位置有关，因此引入了段落的位置信息，即段落在篇章中处于第几段。具体而言，本文对段落位置信息进行编码，得到段落位置信息表示序列$T_{position}=\{e_pos_1,e_pos_2,...,e_pos_5\}$。

将最终的段落语义表示T_{Se}、范畴类别表示T_{e_cls}、位置信息$T_{position}$进行拼接，得到段落的最终表示T_{V_f}，如公式（8）所示。将T_{V_f}送入到softmax层识别出其功能语用，使用$\hat{y}_i$表示段落功能语用的预测结果，如公式（9）所示。

$$T_{V\_f}=concat(T_{Se},T_{e\_cls},T_{position}) \tag{8}$$

$$\hat{y}_i=softmax(T_{V\_f}) \tag{9}$$

本文同样采用负对数似然作为损失函数，记为L_3，如公式（10）所示。其中θ为可训练的参数。

$$L_3(\theta)=-\frac{1}{N}\sum_{i=1}^{N}log(\ \hat{y}_i) \tag{10}$$

4.5 损失函数

本文共有三部分损失，分别为功能语用范畴划分层损失L_1，如公式（2）所示，范畴识别层的损失L_2，如公式（5）所示，和功能语用识别的损失L_3，如公式（10）所示。将最终的损失函数记为L，L计算方法如公式（11）所示：

$$L=L_1+L_2+L_3 \tag{11}$$

5 实验

5.1 实验设置

为了进一步扩大语料规模，依照MCDTB(Jiang et al., 2018)的标注方法，本文在MCDTB的基础上又新标注了480篇与新华社的新闻稿统一风格的新闻报道，形成了规模为1200篇的宏观篇章语料库MCDTB 2.0。本文在宏观篇章语料库MCDTB 2.0上对本文提出的篇章功能语用的识别方法进行了评估。宏观篇章语料库MCDTB 2.0共包含720篇来自宾州篇章树库(Chinese Treebank 8.0,简称CTB 8.0)的新闻报道和480篇来自Gigaword 2.0的新闻报道，共1200篇新闻报道，标注了6763个段落的功能语用，包含15种功能语用类型。

使用Pytorch作为深度学习框架，学习率为1e-4，训练轮数为20轮。文本编码层使用的XLNet版本为XLNet base，最大长度设置为1024，Bi-GRU编码器隐藏层维度设为512,位置信息以及范畴类别编码维度均设置为10，dropout设置为0.1，使用Micro-F1和Macro-F1来分析系统的性能。

5.2 实验结果

为了验证本文提出的模型的有效性，与基准系统进行了对比。各基准系统介绍如下：

（1）SVM：本文复现了Yarlott et al. (2018)的传统机器学习模型SVM来识别篇章功能语用。使用的特征分别为词袋模型、TF-IDF、段落语义特征和上一段的标签信息。

（2）特征+CRF：本文复现了Banisakher et al. (2020)的模型来识别篇章功能语用。除了SVM使用的特征外，还采用了词汇、位置和句法特征，并通过CRF模型来识别篇章功能语用。

（3）Song et al. (2020)：本文复现了Song的序列标注模型识别段落功能语用。此模型将段落语义信息与段落位置信息相结合，并通过自注意力机制获得更加丰富的段落表示，从而识别段落功能语用。

（4）Choubey et al. (2020)：本文复现了Choubey采用序列标注的模型识别段落功能语用。此模型使用分层的Bi-LSTM获得字符、段落和篇章之间的交互信息，并通过分类器识别段落的功能语用。

（5）XLNet(Yang et al., 2019):本文使用XLNet预训练模型获得段落语义表示，并通过分类器识别段落功能语用。

模型	Micro-F1(%)	Macro-F1(%)
SVM	53.23	16.72
特征+CRF	56.28	17.55
Song	64.35	18.27
Choubey	65.13	18.73
XLNet	62.26	15.02
FPRNS(Ours)	**68.19**	**22.63**

表 1. 不同模型的实验结果

实验结果如表1所示。从表1可以看出神经网络模型的识别性能在Micro-F1上均要优于传统机器学习模型，这说明相比传统机器学习模型，神经网络模型能够捕获到深层的语义信息。但是由于传统机器学习模型相比XLNet模型使用了除语义信息之外的结构特征，所以在Macro-F1上的性能高于XLNet模型。

本文提出的FPRNS模型在Micro-F1和Macro-F1的性能均取得最优值，分别达到了68.19%和22.63%，相较于表现最好的Choubey在Micro-F1和Macro-F1上分别取得了3.06%和3.9%的提升。相较于XLNet模型，FPRNS模型在Micro-F1提升了5.93%，在Macro-F1上提升了7.61%。这是由于相比XLNet模型，本文提出的FPRNS模型既捕获到篇章中段落之间交互信息，又融入了新闻图式结构信息，能够获得包含丰富信息的篇章段落表示，与此同时又结合了篇章位置信息，从语义和结构两个方面获得了更加准确的段落表示，因此FPRNS模型对于篇章功能语用的识别能力更强。

5.3 实验分析

5.3.1 消融实验

模型	Micro-F1(%)	Macro-F1(%)
base	65.61	19.00
+position	67.21	21.73
+TS	66.65	21.27
+TS+TSR	67.15	22.09
+TS+TSR+position	**68.19**	**22.63**

表 2. 消融实验结果

消融实验结果如表2所示。base表示先使用XLNet对篇章中的段落进行编码，然后将段落编码信息送入到Bi-GRU中获得具有上下文交互信息的段落语义表示；+position表示在based的基础上加入段落位置信息；+TS表示在base的基础上加上范畴划分信息；+TS+TSR表示在考虑范畴划分信息的同时添加范畴分类信息；+TS+TSR+position表示在融入范畴划分信息的同时结合段落在篇章中的位置信息。

从表2可以看出，加入位置之后，对于篇章功能语用的识别性能在Micro-F1上和Macro-F1上分别有1.6%和2.73%的提升。这是因为段落的位置信息对于位置比较固定的功能语用类型的识别性能影响较大。例如，位于第一段“导语”的识别性能由原来的83.9%提升到了92.31%，提升了8.41%；常位于最后一段对全文做出补充的“补充”类别的识别性能提升了3.8%。

范畴	模型	评价	结果	陈述	补充	次总述	情景	导语	背景	Mic-F1	Mac-F1
总述	base	0	0	0	26.67	19.05	63.55	92.14	21.05	72.98	14.83
	+TS	0	0	0	**30.77**	24.06	65.00	93.04	22.22	73.23	15.66
	+TS+TSR	0	0	0	30.12	**26.61**	**65.21**	**94.57**	**23.34**	**73.51**	**15.99**
故事	base	0	0	18.18	14.52	52.17	80.10	0	0	64.3	11.00
	+TS	0	7.12	19.38	27.74	56.49	80.69	0	8.70	65.87	13.28
	+TS+TSR	0	**7.74**	**20.44**	**30.43**	**58.77**	**80.83**	0	**15.11**	**66.93**	**14.91**
收尾	base	19.05	0	0	70.24	0	27.12	0	0	51.88	9.70
	+TS	27.27	0	0	71.35	0	**28.57**	0	0	53.69	10.70
	+TS+TSR	**34.78**	0	0	**75.35**	0	27.63	0	0	**56.89**	**11.46**

表 3. 各范畴下功能语用识别消融实验

从表2可以看到加入功能语用范畴划分信息后，模型对于功能语用的识别性能在Micro-F1和Macro-F1上均有提升。因为加入范畴划分信息相当于融入了篇章的组织结构信息，能够获得更加准确的段落表示。表3是各范畴下功能语用识别消融实验结果。从表3可以看出加入范畴划分信息之后，对于每个范畴下功能语用的识别性能均有所提升。其中“次总述”、“导语”、“背景”、“补充”、“评价”等功能语用性能的提升更大，因为每个范畴内部内容的组织与整篇文章的组织方式是类似的，即按照“事件总述(导语/次总述)—事件背景介绍(背景)——事件详细阐述(情景)——事件评议(补充/评价)”的形式进行组织。从表2可以看出，加入范畴类别信息之后，对于功能语用的识别性能有小幅提升，同时从表3可以看出每个范畴下功能语用的识别性能也都有所提升，这是由于“导语”仅存在“总述”下，“补充”、“评价”等通常在“收尾”处出现，所以当范畴类别信息作为特征加入之后，能够提升功能语用的识别性能。

5.3.2 错误分类样本分析

功能语用	评价	补充
补充	28.3	/
情景	54.3	39.0

表 4. 错误分类样本的比例(%)

为了分析FPRNS模型的混淆情况，本文统计了错误分类样本，其中混淆比较严重的类别如表4所示。从表4可以看出，因为“情景”类的数量较多，导致有54.3%的“评价”识别为“情景”类。而“补充”和“评价”在篇章位置相似，常位于全篇最后或每个范畴最后，和篇章主体部分语义联系相对松散，并且这两类功能语义相似，因此存在较大混淆。同时有39.0%的“补充”类识别为“情景”类：一方面是因为“情景”类数量较多，另外一方面由于“补充”类在语义上与“情景”类似，所以“补充”类与“情景”类存在混淆。

6 总结

现有的篇章分析研究主要是基于形式语法的，而篇章作为一个语义单位，其功能和语义却没有引起足够的重视，Van Dijk的新闻图式理论指出了段落在整篇文章发挥重要的作用，因此本文提出了一个基于新闻图式结构的篇章功能语用识别方法（FPRNS）识别段落的功能语用。该方法在获取到段落交互信息的同时又融入了篇章的新闻图式结构信息，并结合段落所在篇章中的位置信息，从而有效地提高了篇章功能语用的识别能力。在MCDTB 2.0的实验结果表明，本文提出的方法在Micro-F1和Macro-F1上均取得了最优性能，充分说明了本文提出方法的有效性。由于FPRNS在Macro-F1上的性能还有很大的提升空间，未来将挖掘更加丰富的篇章语义信息以识别出更多样本数量比较少的功能语用类型。本文主要针对新闻领域进行篇章功能语用的识别，此外，像新闻评论、行政裁定书和刑事裁定书等类型的篇章尽管有不同的组织形式，但是却有类似的结构和功能，未来将对这些类型的篇章进行标注并进一步深入研究其结构和功能语用。

第二十一届中国计算语言学大会论文集，第120页-第131页，南昌，中国，2022年10月14日至16日。

参考文献

Deya Banisakher, W Victor Yarlott, Mohammed Aldawsari, Naphtali Rishe, and Mark Finlayson. 2020. Improving the identification of the discourse function of news article paragraphs. In *1st Joint Workshop on Narrative Understanding, Storylines, and Events (NUSE 2020)*.

Allan Bell and Peter Donald Garrett. 1998. *Approaches to media discourse*. Wiley-Blackwell.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Prafulla Kumar Choubey and Ruihong Huang. 2021. Profiling news discourse structure using explicit subtopic structures guided critics. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1594–1605.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Xiaomin Chu. 2018. *Research on representation schema, resource construction and computational modeling of macro discourse structure.* Ph.D. thesis, Soochow University.

Xiaomin Chu, Feng Jiang, Sheng Xu, and Qiaoming Zhu. 2018. Building a macro chinese discourse treebank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Anita de Waard, Paul Buitelaar, and Thomas Eigner. 2009. Identifying the epistemic value of discourse segments in biology texts (project abstract). In *Proceedings of the Eight International Conference on Computational Semantics*, pages 351–354.

MA Halliday. 1994. An introduction to functional grammar 2nd edition, london: Arnold. *Halliday, Michael and Matthiessen, Christian (2004) An Introduction to Functional Grammar, London: Hodder*.

Feng Jiang, YX Fan, XM Chu, PF Li, QM Zhu, and Fang Kong. 2021. Hierarchical macro discourse parsing based on topic segmentation. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, pages 13152–13160.

Feng Jiang, Sheng Xu, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2018. Mcdtb: a macro-level chinese discourse treebank. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3493–3504.

Joost G Kircz. 1991. Rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval. *Journal of documentation*.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Elizabeth DuRoss Liddy. 1991. The discourse-level structure of empirical abstracts: An exploratory study. *Information Processing & Management*, 27(1):55–81.

Yoko Mizuta, Anna Korhonen, Tony Mullen, and Nigel Collier. 2006. Zone analysis in biology articles as a basis for information extraction. *International journal of medical informatics*, 75(6):468–487.

Fairclough Norman. 1995. Media discourse. *London: Edward Arnold*.

Zhongdang Pan and Gerald M Kosicki. 1993. Framing analysis: An approach to news discourse. *Political communication*, 10(1):55–75.

Wei Song, Ziyao Song, Ruiji Fu, Lizhen Liu, Miaomiao Cheng, and Ting Liu. 2020. Discourse self-attention for discourse element identification in argumentative student essays. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2820–2830.

Zhenhua Sun, Feng Jiang, Peifeng Li, and Qiaoming Zhu. 2020. Macro discourse relation recognition via discourse argument pair graph. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 108–119. Springer.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117.

Teun A Van Dijk. 1988. *News as discourse*. University of Groningen.

W John Wilbur, Andrey Rzhetsky, and Hagit Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC bioinformatics*, 7(1):1–10.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

W Victor Yarlott, Cristina Cornelio, Tian Gao, and Mark Finlayson. 2018. Identifying the discourse function of news article paragraphs. In *Proceedings of the Workshop Events and Stories in the News 2018*, pages 25–33.

Xiaoxia Zhang. 2014. The enlightenment and application of macrostructure theory in teaching english newspaper reading. *Journal of Xi'an University of Arts and Science(Social Sciences Edition)*, 17(1):81–84.