CCL23-Eval 任务5总结报告：跨领域句子级别中文省略消解

李炜，邵艳秋*，祁佳璐
北京语言大学
信息科学学院
北京市海淀区学院路15号，100083
liweitj47@blcu.edu.cn, yqshao163@163.com, jialuqi983@163.com

摘要

省略是一种会出现在包括中文在内的各种语言中的一种语言现象。虽然人类一般能够正确理解带有省略的文本，但是其对机器在句法、语义等方面的理解却会造成影响。因此自动恢复省略成分对文本自动分析理解具有重要意义。本任务提出一个面向应用的省略恢复任务，旨在恢复在句子句法结构中占据有效位置同时在句子中扮演语义成分的被省略内容。本任务将省略恢复任务划分成两个子任务：省略位置探测和省略内容生成，并分别描述在两个子任务中取得较好结果的基线方法。此外，为了推进对大语言模型的研究，本文还尝试使用场景学习的方法使用ChatGPT来完成本任务，并进行了相关分析。

关键词： 省略探测 ；省略恢复

Overview of CCL23-Eval Task 5: Sentence Level Multi-domain Chinese Ellipsis Resolution

Wei Li, Yanqiu Shao, Jialu Qi
Beijing Language and Culture University
School of Information Science
15 Xueyuan Rd., HaiDian District,
Beijing, 100083
liweitj47@blcu.edu.cn, yqshao163@163.com, jialuqi983@163.com

Abstract

Ellipsis is a linguistic phenomenon that occurs in various languages, including Chinese. Although humans can generally understand text with omissions correctly, it can have an impact on machine understanding in terms of syntax and semantics. Therefore, the automatic recovery of omitted elements is of significant importance for automated text analysis and comprehension. This task proposes a computationally feasible omission recovery task that aims to restore omitted constituents that occupy valid positions in the syntactic structure of a sentence while playing a semantic role. The task is divided into two subtasks: ellipsis position detection and ellipsis content generation. Baseline methods that have achieved good results in both subtasks are described. Additionally, to advance research on large language models, this study also attempts to utilize the approach of in context learning using ChatGPT to perform this task and conducts relevant analysis.

Keywords: Ellipsis Detection , Ellipsis Restoration

* 通讯作者 Corresponding Author

项目资助：本成果受国家自然科学基金项目（61872402）,教育部人文社科规划基金项目(17YJAZH068)，北京语言大学校级项目（中央高校基本科研业务费专项资金）（22YJ080002）资助。

text	貌似今天我很忙啊，从上午直到现在都在机房维修服务器，现在恢复正常了
targets	index: 9, content: 我 index: 28, content: 服务器

Table 1: 省略任务样例

1 引言

省略是一种在包括中文在内的各种语言中均广泛出现的常见语言现象。虽然省略的相关信息未在文本中出现，但是人类通常可以通过上下文推断出被省略的成分。然而对于机器来说，省略现象却可能导致理解的错误，比如语法分析的错误、机器翻译中的错漏等现象(吕叔湘, 1990)，这使得自动探测和补全省略现象十分重要。前人对于省略现象的研究大多从语言学角度出发。黎锦熙(黎锦熙, 2007)提出省略是担任主语、宾语等句法成分的实体词的省略，吕叔湘(吕叔湘, 1990)等人沿革并发展前任思想，至王维贤提出“三个平面”理论，从语义、语法、语用三个角度深入论述省略现象(王维贤, 1985)。八十年代以来，吕叔湘(吕叔湘, 1979)、祝克懿(祝克懿, 1987)、陈平(陈平, 1987)等人进一步对比区分省略与零形回指、隐含、暗示等省略现象，拓宽了对省略的探索和研究。简单来说，省略可以被分成三个类别：句法层面、语义层面和语用层面。前人研究较多的零指代消解可以被归为句法层面省略的一种。Ren等人(Ren et al., 2018)构建了针对于网络文本的省略恢复数据，但是其中恢复的省略内容可能超出当前句子。而Liu等人(Liu et al., 2019a)则是基于AMR图构造省略数据集，其省略的成分可以以符号的形式进行补充。

在本任务中，我们主要从应用和可计算的角度出发，来定义省略的范畴和成分。我们将省略看做是未出现成分在句子句法结构中应该占据有效位置，同时在句子中扮演语义成分，并且能够根据句子内上下文信息补充的句子成分。这样使得被省略成分能够依据有限的上下文（限定在句子内，不需要额外背景）通过计算被还原出来，同时对被省略成分进行还原后，能够提升句法结构和语义结构的完整性，帮助计算模型对句子的理解。

比如例子“**妈，为什么一听见他的声音，我的双腿就发抖**”中，就省略了主语“**我**”，也即是“**我**”听见声音。（完整句子应为“**妈，为什么我一听见他的声音，我的双腿就发抖**”）

被省略的成分“我”可以从后文中找到，同时使得中间小句的句法结构比较完整，并且得到了是我听声音这样的语义信息。

除了此类与零指代消解有一定重合的情况，句子中的其它成分也可能会被省略，比如“你吃饭了吗？我没。”该例中就省略了谓语成分“吃”。而对于无法单纯从句子内信息进行补全的成分，本任务不对这类省略进行研究。比如“中国从前的监狱，墙上大抵画着一只虎头，所以叫做‘虎头牢’，狱门就建在虎口里，这是说，一进去，就很难再出来”这里就省略了需要通过知识背景才能补全的“犯人”或“罪犯”。

在任务中，我们整理发布了经人工标注的5953条包含有本文定义的可恢复省略成分的句子，同时标注了省略的位置和省略的内容。出于比赛的需要，我们将数据集随机划分成训练集、开发集、测试集和盲测集四个部分（见表3）。此外，我们观察到在不同的语体中，省略现象出现的概率和类型等会有差别，因此我们同时选择了相对正式的新闻、教材、小说和剧本以及相对非正式的微博和产品评论数据，使得数据更加具有广泛性和实用性。

在本任务中，我们将任务具体划分成两个子任务（见表1）：省略位置探测和省略内容恢复。省略位置探测旨在在明确句子中存在省略的情况下，找出省略的具体位置（下标），也即找到标注中的”index”，在此例中为9和28，因为省略位置应为“从”和“恢复”前面的位置，而“从”从0开始计数为9，“恢”从0开始计数为28。在本任务的基线模型中采用序列标注的形式进行建模。省略内容恢复旨在在已知省略位置的情况下，找出或生成出被省略的内容，在此例中为“我”和“服务器”。在本任务的基线模型中采用文本生成或问答的形式进行建模。此外，本任务还包含将以上两个子任务一体化的总任务，即已知句子中存在省略，需要同时判断省略的位置并给出被省略的内容。目前的一体化总任务的基线模型为前述两个子任务获得最佳性能的流水线方法。

本文的贡献可以总结如下：

- 我们发布了一个有5953个来自6个不同领域的标注了省略位置和省略内容的句子的省略恢复

第二十二届中国计算语言学大会论文集，第159页-第166页，哈尔滨，中国，2023年8月3日至5日。

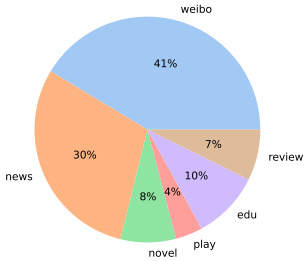

Figure 1: 不同领域样本分布比例

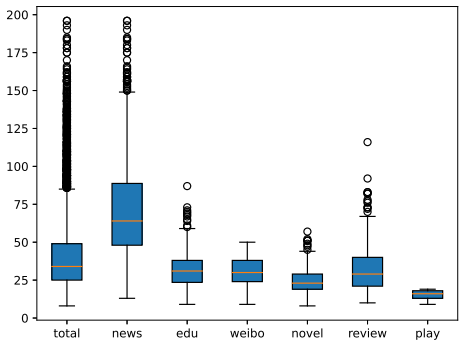

Figure 2: 不同领域样本文本长度分布箱线图

任务。通过省略恢复任务，本研究希望能够提高句法和语义分析的准确率，以更好地分析和理解自然语言文本。

- 我们发布了包括当下获得巨大成功的大模型ChatGPT[0]在内的多个省略恢复基线模型。基于ChatGPT的方法通过不同的提示语，以少样本学习的方式测试大模型对省略现象的恢复能力。本研究希望能够为大模型生成内容与省略现象的进一步研究提供基础。

2 数据集

我们选取的数据包含相对正式的新闻、教材、小说、剧本，以及相对非正式的微博和产品评论。相比MCER(Qi et al., 2022)的数据，主要增加了口语化的微博数据，使得研究更加贴近当前网络文本大量出现的情况。其中，为方便后续从语义依存分析的角度对省略补全的意义进行验证，新闻、教材、小说、剧本四类数据的语料来源均与CCL2020中文语义依存图分析任务评测中给出数据集的语料来源相同。我们以句子通顺、无歧义，且包含一个或一个以上省略现象为原则，最终得到共5953句带有省略位置和省略内容的标注。

各领域具体句子数量和比例如表2和图1所示。可以看到数据集中以微博和新闻所占比例最高，分别代表了正式文本和非正式文本。我们还在图2中展示了不同领域样本的文本长度分布，可以看到总体长度较短，但是也存在大量较长样本，这部分主要来自新闻领域。而受到平台的限制，微博文本则普遍较短且基本没有异常值。

微博	新闻	小说	剧本	教材	产品评论
2463	1774	471	232	579	434

Table 2: 不同领域样例数量

训练集	开发集	测试集	盲测集
3,606	602	883	862

Table 3: 数据集划分

在表3中我们展示了比赛中数据被如何划分成不同部分。需要说明的是，最终的比赛排名按照盲测集结果排名。

在图3中我们展示了不同领域中样本省略内容长度的分布，可以看出省略长度大部分在5以内，但是产品评论省略内容总体长度偏长，这可能与产品评论中提到的被省略内容多为产品的某些方面，而这些描述往往偏长有关。小说和剧本的省略长度相对较短，并且很少有异常值，这可能与文学作品中省略内容多为其中人物，而人物名长度一般较短有关。

在图4中我们展示了不同领域中省略个数的分布，可以看出省略个数大多均为一到两个，且不同领域差别不是很大。

2.1 数据标注过程

我们的数据标注过程主要分为人工标注和二次验证两个部分。

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

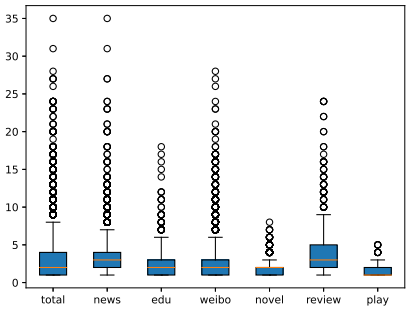

Figure 3: 不同领域样本省略内容长度分布箱线图

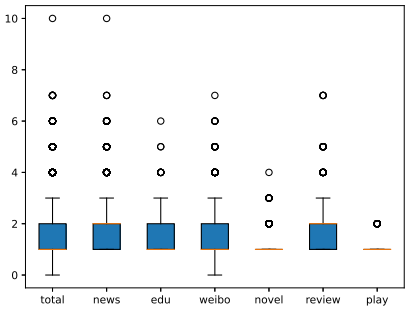

Figure 4: 不同领域样本省略数量分布比例箱线图

人工标注阶段分为试标注和正式标注两个步骤。我们通过试标注过程中标注人员的反馈来确认标注规范中是否存在歧义或难以理解的内容，并依据实际情况，对标注规范进行调整。与此同时，我们也会对标注人员试标注过程的标注结果进行检查，针对其结果中与正确的标注结果存在差异的部分，与标注人员沟通，并借助这些问题以及一些典型案例，对标注人员进行进一步的培训，确保其能够充分理解标注规范的内容及其含义。正式标注过程中，标注人员两两一组，每组标注人员将对同一组数据进行标注，以方便我们进行后续的二次验证。

在二次验证阶段，我们对每组的两个标注人员提交的标注结果进行比对，并针对标注结果中存在差异的部分，进行重新标注，直至两个标注人员的标注结果达成一致为止。

3 任务

在本节中，我们对两个子任务以及在两个子任务取得较好效果的基线模型进行描述。

3.1 子任务一

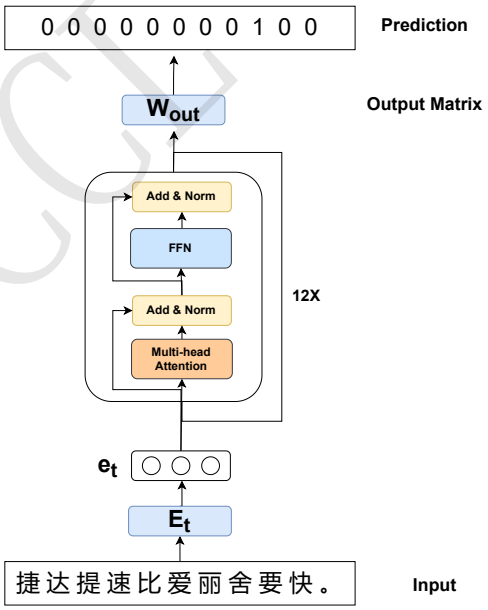

Figure 5: 序列标注方式进行省略位置探测方法示意图

对于子任务一，任务的输入为带有省略现象的文本原句，如图5中输入为“捷达提速比爱丽

舍要快”，这里完整的句子应为“捷达提速比爱丽舍**提速**要快”，省略内容为“提速”，其被省略位置应为“要”字前，因此本任务输出答案设置为“要”字在当前句子中的下标，也即从0开始计数的第8个位置。

在此子任务中，本文选取的基线模型将此任务建模成序列标注任务，也即对于输入文本的每个位置进行角色预测，省略所在位置标注为1，其它位置标注为0。在例子中，“要”字对应的序列目标为1，其余位置为0。如图5中输出目标所示。本任务采用的实现方式为Transformers[1]。对于编码器部分来说，本文分别选取了BERT(Devlin et al., 2019)和Roberta(Liu et al., 2019b)。输出矩阵将经过编码器编码的每个字符对应的隐向量转换到二分类概率。并使用交叉熵作为优化目标函数。

3.2 子任务二

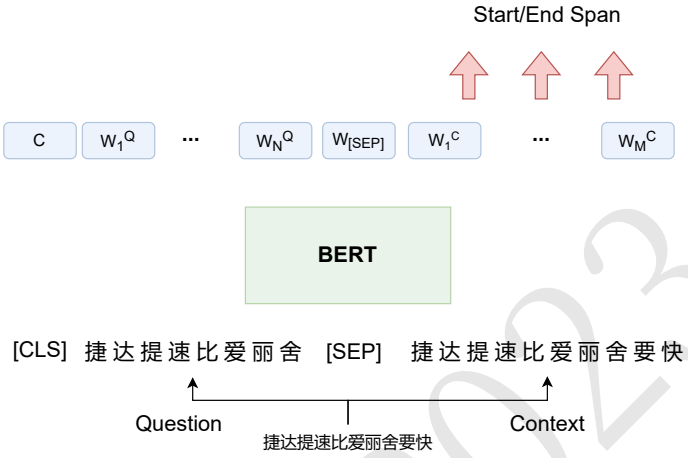

Figure 6: 问答方式进行省略内容补全方法示意图

对于子任务二，任务的输入仍为带有省略现象的文本原句，标准答案输出应为被省略内容“提速”。在此子任务中，可以直接使用序列到序列的建模方式，即以原任务输入为模型输入，以被省略内容“提速”为模型输出(Lewis et al., 2020)。但是此方法的实际效果较差，因此本文着重介绍使用阅读理解形式建模的方法。阅读理解形式方法将输入重新构造成问题和上下文两个部分（如图6所示），使用[SEP]隔开。输入句首插入[CLS]代表文本整体语义信息。基于该位置得到的整体表示与上下文信息进行比较，分别获得答案开始和结束的位置，并使用交叉熵作为优化目标。

4 结果

4.1 评价指标

对于子任务一来说，本任务选择精确率（Precision）、召回率（Recall）和F1值进行评价，我们以F1值作为子任务一的主要评价指标。对于子任务二来说，考虑到生成的内容可能并不完全重合，在完全匹配之外，本任务还选择了Rouge-1、Rouge-2、Rouge-L对生成内容进行评价。子任务二的最终评价分数由精确匹配分数和Rouge分数加权求和得到：

$$score_2 = 0.4 * Exact\text{-}match + 0.3 * Rouge\text{-}L + 0.2 * Rouge\text{-}2 + 0.1 * Rouge\text{-}1 \quad (1)$$

一体化任务评价方法与子任务二评价方法保持一致。

4.2 大语言模型

为了探究当前在各种不同任务中取得良好效果的大语言模型使用情境学习（in context learning）方法在省略位置探测和省略内容恢复任务上取得的效果，本文设计了两段提示语来分

[1] https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertForTokenClassification

别使用尖号∧标记省略位置（表4）和填充[$MASK$]对应位置的被省略内容（表5）。在提示语中，我们首先通过文本描述任务目标，之后提供三个包含有输入、输出对的例子，并在文本的最后提供带判断的文本，指示模型在“输出”后参照例子给出预测结果。

省略指句子中对语法或语义成分的省略。请参考以下例子，以∧标注省略现象出现的位置。
输入：他俩并不在看电视，只是借电视来营造一个只属于他俩的氛围，以这氛围在这家中做一种微妙的划分。
输出：他俩并不在看电视，∧只是借电视来营造一个只属于他俩的氛围，∧以这氛围在这家中做一种微妙的划分。
输入：看着一拨一拨的人陆续上车，我们的车还没来...
输出：∧看着一拨一拨的人陆续上车，我们的车还没来...
输入：捷达提速比爱丽舍要快。
输出：捷达提速比爱丽舍∧要快。
输入：sentence
输出：

Table 4: 省略位置探测ChatGPT提示语

省略指句子中对语法或语义成分的省略。以下示例中[MASK]表示这一位置为省略出现的位置，请参考这些示例，在句子内部找到合适的成分对替换[MASK]的部分，注意用于替换的成分必须在句子里出现过。
输入：他俩并不在看电视，[MASK]只是借电视来营造一个只属于他俩的氛围，[MASK]以这氛围在这家中做一种微妙的划分。
输出：他俩他俩
输入：[MASK]看着一拨一拨的人陆续上车，我们的车还没来...
输出：我们
输入：捷达提速比爱丽舍[MASK]要快。
输出：提速
输入：masked_text
输出：

Table 5: 省略内容补全ChatGPT提示语

4.3 参赛队伍结果

在表6中，我们展示了提交结果的两支队伍和基线方法以及大语言模型ChatGPT在子任务一中得到的结果。其中，基线方法使用BERT作为语义表示模型。

可以看到两支队伍在子任务一中的结果并不理想，未达到本文基线方法的水平。而相比两支队伍来说，ChatGPT取得的结果则更差。通过对ChatGPT结果的简单分析，我们发现，在盲测集的862个样例中，有711个样例ChatGPT都发生了不同程度的错误。其中，甚至在高达465个样例中，省略的数量与标准答案不匹配。

队伍	Precision	Recall	F1-score
Baseline	81.29	81.11	81.20
ChatGPT	39.99	46.49	42.99
北京语言大学	59.55	70.06	64.38
大连理工大学	67.01	75.99	71.22

Table 6: 子任务一参赛队伍盲测结果

在图7中，我们给出了基线模型在不同领域中在子任务一中得到的F1分数，可以看到尽管是正式文体，新闻领域中的省略位置探测与微博类似，均低于其它领域。我们认为这可能与新

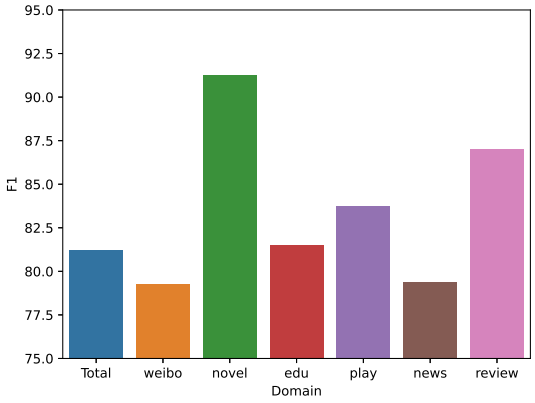

Figure 7: 不同领域下子任务一中基线模型得到的F1分数

闻文本长度相对较长有关（见图2），而微博中的省略位置探测结果差则主要和其表达更加随意有关。而由于在整个数据集中新闻和微博所占比例较高，因此总的F1分数也与这两个领域的分数较为接近。

在表7中，我们展示了提交结果的两支队伍在一体化任务中得到的结果。可以看到大连理工大学队获得的成绩相对于北京语言大学队的成绩更优，但是两支队伍综合得分尚未达到本文发布的基线模型。而目前使用的基于ChatGPT的模型在一体化任务中结果与两支队伍差距较大，说明基于ChatGPT的省略恢复仍需要继续进行研究。

队伍	Rouge-1	Rouge-2	Rouge-L	Exact Match	综合得分
ChatGPT	20.02	12.21	19.99	15.15	16.50
Baseline	69.27	46.91	69.27	64.20	62.77
北京语言大学	35.96	20.15	35.91	23.94	35.39
大连理工大学	55.05	30.80	55.04	48.17	55.37

Table 7: 一体化任务参赛队伍盲测结果

由于盲测集的设定，参赛队伍没有单独子任务二的结果（子任务一中的省略位置给出的情况下），因此我们仅在表8中给出基线模型和ChatGPT在子任务二中的表现。可以看出，虽然ChatGPT是生成模型，但是在此类限定范围的生成任务中，与经过训练的问答模型表现上仍有较大差距。

方法	Rouge-1	Rouge-2	Rouge-L	Exact Match	综合得分
Baseline	82.32	54.98	82.29	75.15	73.98
ChatGPT	39.61	22.03	39.57	30.16	30.30

Table 8: 子任务二基线方法结果

5 结论和展望

在本文中，我们介绍了跨领域句子级别中文省略消解恢复评测的任务设定和本文所采用的基线方法，以及参赛队伍的结果。从结果来看，参赛队伍在两个子任务上获得的结果均尚有较大提高空间。此外，为了探究大语言模型对省略现象的理解能力，本文尝试使用场景学习的方法使用ChatGPT来完成本任务。

在未来，本文认为可以对语义分析与省略恢复之间的关系进行研究，比如省略恢复是否以及如何能够帮助语义分析取得更好的效果。此外，本文对大语言模型在省略现象上的研究还非常不充分，在未来的研究中，我们认为可以探究大语言模型所生成内容中带有的省略现象是否与人类生成文本中带有的省略现象一致等方面。

参考文献

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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)*, pages 4171–4186. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Yihuan Liu, Bin Li, Peiyi Yan, Li Song, and Weiguang Qu. 2019a. Ellipsis in Chinese AMR corpus. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 92–99, Florence, Italy, August. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jialu Qi, Yanqiu Shao, Wei Li, and Zizhuo Shen. 2022. Mcer: A multi-domain dataset for sentence-level chinese ellipsis resolution. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 29–42. Springer.

Xuancheng Ren, Xu Sun, Ji Wen, Bingzhen Wei, Weidong Zhan, and Zhiyuan Zhang. 2018. Building an ellipsis-aware Chinese dependency treebank for web text. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

吕叔湘. 1979. 汉语语法分析问题. 商务印书馆.

吕叔湘. 1990. 中国文法要略. 商务印书馆.

王维贤. 1985. 说“省略”. 中国语文, 6:409–414.

祝克懿. 1987. 省略与隐含. 河南大学学报*(*哲学社会科学版*)*, (05):92–97.

陈平. 1987. 汉语零形回指的话语分析. 中国语文, 1987:363–378.

黎锦熙. 2007. 新著国语文法. 湖南教育出版社.