中文谜语任务-作业报告

北京大学 2021-2022 秋季《互联网数据挖掘》

小组成员: 李思哲、戴舒羽、侯树頔
学生学号: 1900013061、1900013074、1900012987
学院: 北京大学 信息科学技术学院
指导教师: 万小军

1 任务描述

在这个任务中，我们的目标是训练一个能够解答中文谜语的网络。

与传统的 NLP 问答任务不同，谜语往往需要更多地了解答案本身，需要模型具有一定的知识储备，从而可以解析问题中对于答案某些特征的描述，例如问题 ***If you take off my skin, I will not cry, but you will. What am I?*** 的答案为 ***onion***，这不仅需要模型理解问题的语义，更需要模型了解答案的特征。

而与英文的 riddle 不同的一点是，中文谜语往往还会考虑汉字的字音、字形等信息，例如问题**兀**，答案为**走西口**，其中则考虑到了汉语的字形，即将汉字西中的一部分口去掉，便能获得兀字。因此又为汉语谜语增添了新的难度。

在这个任务中，我们将问题简化，由谜语本身的填空题改为选择题，即给出五个备选答案，训练模型在五个选项中找出正确答案即可。给出的数据包括训练集、验证集、测试集、以及词语的解释。其中训练集和验证集包含问题、选项及答案，测试集只包含问题及选项。

2 完成过程

2.1 文献调研

对于解答中文谜语的任务，我们尚未找到贴切任务的论文，但适当进行外延，我们调研了解决 riddle 任务、以及带有 information 的 QA 任务的方法，调研的文章如下：

1. **Unified QA** Khashabi et al. (2020)

   文章提出了一个简洁的统一各种形式问答任务的模型 UnifiedQA。针对输入，首先将不同问题处理成统一格式（将问题、文段、备选项依次输入模型中，用回车作为分隔符）。在训练过程中，混合使用不同形式的问答任务。在训练中，一条数据被选为训练数据的概率为 $\frac{1}{T_i}$,T_i 代表 i 任务的数据条数，这样平均起来能够保证每个 batch 中来自各任务的数据均匀分布，不受到每个任务数据集规模的影响。

2. **RiddleSense**(Lin et al. (2021))

   文章提出了基于英文 riddle 的数据集 RiddleSense，与我们的中文数据集在形式上极其相似。解决 RiddleSense 问题需要模型依据 commonsense 进行推理，是一种进阶的 Natural Language Understanding 任务。对于如何解决这类问题，文章提出了三种常用的方式，分别为 fine-tuning 现有的 LM；使用符号化的知识图谱进行推理；以及 fin-tuning 现有的 T5 以进行一个 text2text 的训练。文章给出了最后一种方式的一个实现代码。

3. **QAGNN**(Yasunaga et al. (2021))

   文章主要对现有结合知识图谱、语言模型二者的方法进行两点改进：其一，拼接节点和问答文本，输入预训练模型计算知识图谱中节点的权重，筛选知识图谱中对问答最有帮助的节点。其二，将问答文本本身作为一个新节点加入图结构中，使得训练一体化。在构造子图的过程中，先提取问题题面和选项的 entity，再找到所有与这些 entity 存在长度不大于 2 的通路的知识图谱中的节点，加入子图中。之后通过预训练模型计算问答文本的表示，也作为一个新节点加入子图。除了知识图谱中原有的边，增加 QuestionEntity 和 ChoiceEntity 两种边，链接问答文本节点和文本和答案中包含的 entity 节点。具体方法如图1所示。

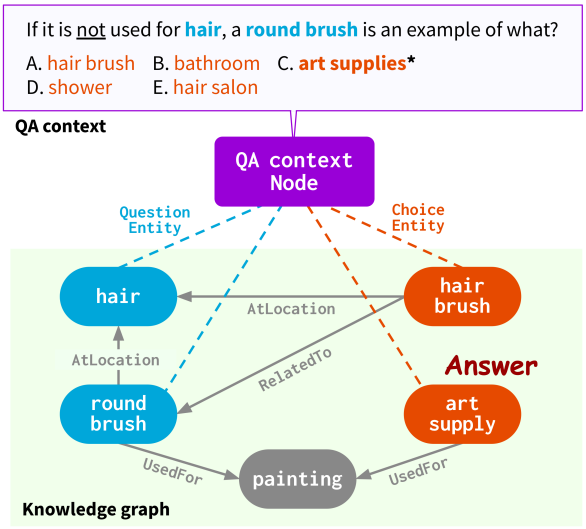

Figure 1: QAGNN 的原理图

2.2 思路

对于解决中文谜语问题，我们的思路如下：

1. 我们发现谜语利用了许多字形的信息，如进行拆字、提取汉字的偏旁部首进行组合等（具体见附录6），相反，谜语对于答案的语义信息的使用相对较少，尤其当答案的解释并不常见时。而对于我们目前的数据集，对答案的解析并不能很好地与谜面相匹配。
2. 按照上述分析，我们认为在正常的分类任务的 pipeline 中使用选项的释义信息不能取得很好的效果，因此我们想要通过知识图谱来构建选项之间的释义的关系。
3. 与从多个选项中选择一个相比，我们认为训练模型做二分类更加简单，尽管需要更多地数据处理工作，但我们将问题简化成了模型只需要判断当前地谜面和选项是否匹配，极大地简化了问题
4. 谜面中除了问题外，还给出了答案的类别，若加以利用，可以为模型提供有效的提示。例如，当将谜语问题视为填空题时，类别信息可以很好地指导模型预测答案。
5. 除此之外，谜语与 QA 不同，考察的主要是跳出原有思维进行“脑筋急转弯”的能力，有时谜面和答案在字面意义上的相似性可能很有限。但正确选项和谜面本身并不是毫无关系，它们也存在着逻辑上和常识上很强的相关性，可以利用这一点选出与谜面最相似的答案。

2.3 几个 baseline

在这里列举尝试阶段我们使用的不同 baselines，其性能会在性能分析章节进行对比。

Similarity 方法 考虑到正确选项和谜面本身也具有较大相似性，我们在这里采用了一种**非训练**的方法，即使用 sentence-transformerReimers & Gurevych (2019) 分别提取谜面和选项的 sentence embedding，分别计算谜面和每个选项的相似度，选择相似度最高的选项作为预测的答案。在这里我们分别尝试了使用选项本身以及选项的信息作为输入。

字形方法 考虑到不少中文谜语都利用了字形信息，我们采用一种**非训练**的方法。利用中文拆字表，将除去括号中的补充信息的谜面 riddle 以及每个选项 choice 逐字拆分成多个小的组成部分（拆字着重于尽量把每个字拆成两个以上的组成部件，而不是手写时使用的笔画），同时保留原有所有单字并加大其权重，然后合成一个 riddle_list 和四个 choice_list。统计 riddle_list 中每一项在各个 choice_list 中出现的次数并加和，所有选项中次数最多者为选择结果。

填空 MLM 方法 考虑到谜语任务的谜面形如“题干（打一事物类别）”，关于“事物类别”的提示大大减少了可能选项的数量，并且预训练模型经过 MLM 预训练任务的训练，可以保证填空内容在语义上的通顺。那么，我们不妨将这个多选任务转化为填空任务。以“小时青青似野草，老来满头金珠宝，珠宝人人都珍惜，天下无人不依靠。（打一植物）”为例，具体如下：

首先将谜面转化为“小时青青似野草，老来满头金珠宝，珠宝人人都珍惜，天下无人不依靠。这个植物是”，加上正确选项“小麦”，构造出正确答案“小时青青似野草，老来满头金珠宝，珠宝人人都珍惜，天下无人不依靠。这个植物是小麦”。构造输入“小时青青似野草，老来满头金珠宝，珠宝人人都珍惜，天下无人不依靠。这个植物是 [MASK][MASK]”，将模型在所有 [MASK] 位置的输出提取出来，假设为“麦草”。之后使用 sentence-transformer，计算出“麦草”与“小麦”和其他备选项的 sentence embedding，将“麦草”的 embedding 和选项的 embedding 分别计算余弦相似度，选择和“麦草”余弦相似度最高的一个选项作为最终答案。

二分类方法 我们前面提到可以将问题归约到二分类问题，只需要将谜面与每个选项组合，并判断其是否匹配。整体的实现可以使用 bert 的 fine-tuned 算法，为每一个选项进行匹配程度的打分，最后取得分最高的选项即可。在这里我们尝试了多种不同的预训练模型，以及是否使用 info 信息。

多选一任务 使用 huggingface 提供的 BertForMultiChoice 接口，首先将数据集.csv 文件处理为特定格式：sentence1-sentence2 为“谜面” - “打某某是（选项）”，每组五个句子对，找出衔接最通顺的句子即为选项答案。尝试在多种预训练模型上 fine-tuned。

RiddleSense 方法 我们使用了论文Lin et al. (2021) 提供的开源代码，首先复现 RiddleSense 的工作。之后将数据处理为 json 格式读入，并尝试使用多种预训练模型进行 fine-tune。

QAGNN 方法 [未完成] 我们使用了论文Yasunaga et al. (2021) 提供的开源代码，首先将 QAGNN 的数据预处理移植到中文。包括筛选中文节点，更改图中节点的 entity embedding 等，并更改预训练模型为可以接受中文的 xlm-roberta 模型

3 实现方法

3.1 谜面分类

我们注意到，存在部分谜面只包含单个字，而这部分谜面提供的语义信息极少，因此我们可以认为这部分谜语完全基于字形进行猜测。因此我们在进行后续操作前，首先判断是否为基于字形的谜语，若是，则直接使用上述的字形方法进行预测。

3.2 Major Voting

在排除了基于字形的谜语后，我们将其余的谜语通过上述两种性能较高的方法（多选一、RiddleSense）分别得到预测，之后利用 Major Voting 算法处理预测结果，选择得到了更多模型确认的选项，或在较好的模型中获得较高得分的选项作为最终预测。

4 结果分析

4.1 性能分析

对于上述的 baseline 和 final method，我们在不同条件下进行了测试，其性能如下表1

	bert-base-chinese		xlm-roberta-base		ernie 1.0		no		单字
	√	×	√	×	√	×	√	×	×
Similarity	-	-	-	-	-	-	28%	32%	-
字形	-	-	-	-	-	-		29%	69%
填空 MLM	-	40%	-	21.3%	-	27.8	-	-	-
二分类			12%		50%	56.4%	-	-	-
多选一		52%		30.4%		62.9%	-	-	-
RiddleSense		48%		22%		**63.4**%	-	-	-
Final		52%	12%	30.4%	50%	**63.4**%	28%	32%	69%

Table 1: 现有方法在验证集上的表现，第一行表示方法使用的预训练模型，单字表示在训练集和验证集所有的单字谜面上的测试结果，第二行表示方法是否使用了选项的 info 信息。标记**为-的表示当前实验设定下不存在该数据，空白表示时间原因尚未得到的数据。**

4.2 预测结果分析

我们认为，由于中文字谜涉及到音形义等多方面特征的综合，因此谜面的长度本身也是一个不可忽视的指标。我们对比了在不同长度（单个字、短语、句子）上不同 baseline 的表现，具体表现如下图2

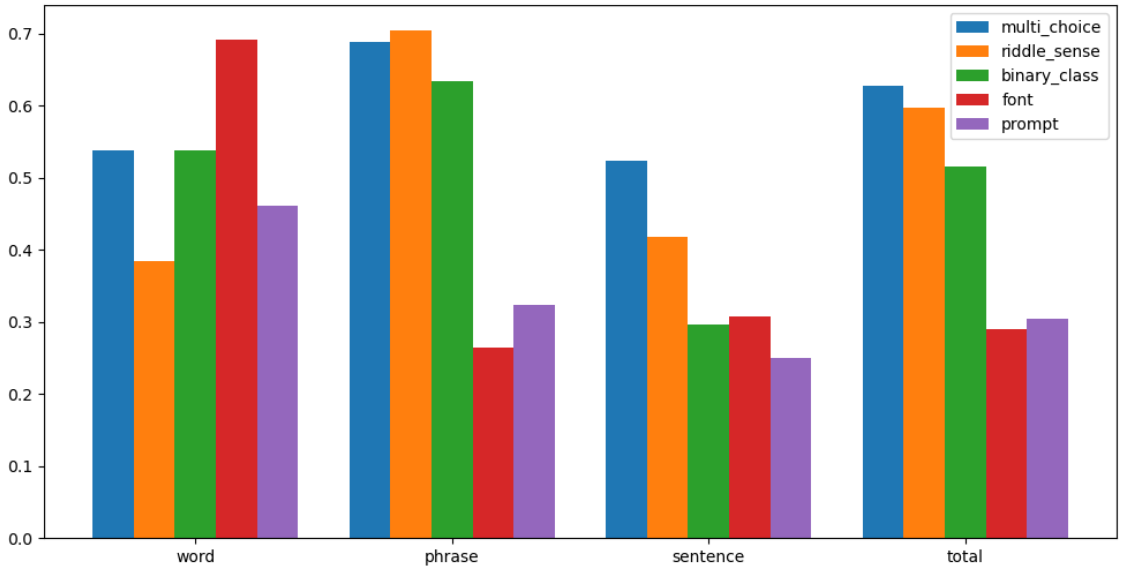

Figure 2: 几种 baseline 在不同长度谜面上的表现的对比

我们发现，除了字形方法（即图中的 font）在单个字的谜面上表现最好之外，其余方法均在谜面是短语时表现最佳。我们认为这是合理的，若谜面仅有单个字，则很大程度不含有语义信息，因此只考虑其字形信息是明智的选择。对于谜面为整句话的情况，由于训练数据较少，且谜面可能较为复杂，需要推理，因此模型表现不如谜面为短语的情况。

此外，对于字形方法，对于单字也依然存在错误情况，例如对于谜面**十（打一成语）**，模型不能正确给出答案**纵横交错**。分析原因在于尽管谜面只有单字，但答案中包含语义信息，但目前的模型并不能很好地处理这种情况。

对于我们最后采用的方法，我们给出验证集上的一个较为典型的错误案例，题面为**一百乌龟一百鳖，爬山过岭勿会跌。（打一物）**，答案为**水车**，模型预测的结果为**碾子**。我们认为这可以代表一类错误，即需要对谜面的每个词语理解，并了解**乌龟**的习性特征，从而进行复杂推理。这一过程对于人类也尚有一定难度。此外，由于训练集规模较小，从而训练得到的模型的泛化能力也还有待提升。

5 未来工作

由于时间仓促以及任务本身的难度之大，我们的方法还尚有许多不足。我们希望能够在学期结束后进一步研究该问题，目前有以下几个想法：

1. 进一步挖掘字形方面的信息

   上述方法中，对于字形信息的使用还仅仅停留在表面层次，未来我们希望能够进一步思考如何选取最可靠的偏旁部首，如何计算偏旁部首的 embedding，以及将字形信息的 embedding 与 bert 的语义信息融合

2. 有效利用数据集提供的 wiki 文档

   我们能够想到的对于 wiki 的最佳利用方式为 GNN。许多包含生僻字的词语、成语、一些古代作品名没有被知识图谱收录（例如：“鹘鹬”、“吮痈舐痔”、“留侯论”），不能使用知识图谱丰富的常识信息，这是十分可惜的。但是 wiki 的提供无疑弥补了这一缺憾。目前不太成熟的想法是在现有的图结构上增加 wiki 信息中包含的节点，补充知识图谱所缺失的信息。

3. 多进行一些模型之间的比较

4. 结合谜面的读音

5. 细化填空题设问

   我们目前的模板只有“这个 XX 是”一种，如果可以利用中文量词丰富的特征，做到根据“XX”的类别对设问进行修改（例如：“这个植物是”修改为“这株植物是”），无疑会给模型更多的信息用于判断。

6. 实验 QAGNN 方法

   由于 QAGNN 需要对很多节点一一测试相关度，代码完整运行需要时间过长，因此没能在规定时间内完成运行。在英语到中文的移植过程中也有很多粗糙之处，希望可以在学期结束之后进行完善。

6 小组分工

李思哲 参与文献调研工作，主要研究文章 RiddleSense 并完成代码的复现工作；完成 RiddleSense 的数据预处理工作；完成 Similarity 方法、二分类方法、RiddleSense 方法，三种 baseline 的代码编写；完成最后的结果整理以及 Major Voting 算法实现；撰写报告。

戴舒羽 调研字形方法，字形方法、多选一任务方法两种代码实现；最后参与结果整理工作

侯树頔 与李思哲、戴舒羽一起讨论确定使用拆字、二分类、填空、知识图谱为主要方案。参与文献调研工作，主要研究文章 QAGNN(Yasunaga et al. (2021)) 并基本完成代码在中文中的移植工作（绝大部分数据处理工作已经完成，随其他模型代码一同附上）；完成填空方法代码编写

附录：提取汉字的字形信息

想法提出 谜语猜题方法中有是不少从字形意义上寻找线索，具体可以归类到以下三种情况：

- 离合法：铆-金兔
- 半面法：扣-各执一端
- 减损法：皇-白玉无瑕　兀-走西口

可行性分析 拆字的作用是提取出一些字形特征，但汉字本身就是象形文字，所以拆字并不应该是简单的比较相同部分，而是可以提取出偏旁部首的语义信息，比如“聋”和“充耳不闻”具有某些相似的语义。

另外，能够 work 的一个很大原因是针对于任务的形式本身——这是选择题，而且干扰项主要是语义上面的，在字形上正确选项反而显得非常突出。谜底中语义提示部分的信息被忽略，比如“各执一端”暗示各取一部分的语义，以及“白玉无瑕”暗示无瑕-> 去掉玉的一点而得到王，但是由于原本的字形局部仍在，所以还是能够依此判断出来。

实验方法 利用 https://github.com/kfcd/chaizi 中的拆字词表，对 riddle 和 choice 进行拆字，然后对结果进行某种相似度比较

存在的不足 单纯通过字形提取信息的占有一定比例，但不适用于每一种情况，比如其他单纯依靠语义，语义结合字形，语义结合语音并不适用。此外，如何融合多方特征进行判断仍然是一个挑战。

References

Khashabi, D., Min, S., Khot, T., et al. 2020, FINDINGS 1

Lin, B. Y., Wu, Z., Yang, Y., Lee, D.-H., & Ren, X. 2021, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP 2021): Findings 1, 3

Reimers, N., & Gurevych, I. 2019, ArXiv, abs/1908.10084 2

Yasunaga, M., Ren, H., Bosselut, A., Liang, P., & Leskovec, J. 2021, ArXiv, abs/2104.06378 1, 3, 5