北京大学

语义计算与知识检索
实验报告

题目：中文字谜问题

组　　员：张文煊 2101213235

刘翠玉 2101213203

徐志杰 2101213200

二〇二二 年 六 月

摘要

本文是北京大学语义计算与知识检索课程的作业，题为中文字谜问题求解。输入一道字谜如“秋后又闻王者香”，算法需给出谜底“烂”。

目前学术界对中文字谜求解问题的研究较少，同时由于语言差异，英文字谜的求解算法也无法完美解决中文字谜问题。因此针对中文字谜的特性，设计一套中文字谜问题求解框架有着重要意义。

本文设计并实现了一套召回加排序的中文字谜问题求解框架，通过对样本的观察统计，发现绝大部分中文字谜可以使用拼字法解决，另外中文的“音形义”要素常常作为拆字拼字的依据。因此召回模型的设计充分考虑了中文字谜的“音形义”特点，再使用拆字拼字法，可以将候选谜底由全体两万余汉字缩减到几十到几百。排序算法使用预训练语言模型，通过训练二分类任务计算预测分进行排序，给出最终预测的谜底。

实验验证了本文提出方法的有效性，在验证集上 MRR@5 指标可达到 24.1%。通过对召回算法的消融实验，可以看出，更全面地考虑“音形义”要素的召回算法能达到更高的召回覆盖率，排序后的结果也更高。

第一章　背景

1.1　问题描述

本次任务的目标是用深度学习模型解答中文谜语。解答谜语属于自然语言处理中的问答任务，但是相对来说更复杂、更难求解，原因主要有以下几点：第一，谜语语料数据待完善，很多谜语人工无法理解或者有多个合理的答案，问题本身存在不确定性，这对模型而言是很大的挑战；第二，求解谜语需要很大的知识储备，需要从多方面提取谜语的特征获得结果，例如问题 *What has a round face and two thin hands, one hand short, one hand long?* 的谜底为 *clock*，这需要模型理解谜语的语义，同时抓住谜底的特征。

现有的工作大多数是解答英文谜语，而且通常是从较少的选项中选取谜底，我们的任务是处理中文字谜，而且要求无选项地输出结果，毫无疑问我们的任务更难处理，除了语义，中文字谜往往还需要考虑汉字的字形、发音等信息，需要对谜语有更全方位的理解，例如字谜半数嫌贵的谜底为赚，即汉字“嫌”和“贵”各取一半得到汉字“赚”，再如字谜走西口的答案为兀，即将汉字“西”中的一部分“口”去掉得到汉字“兀”。

在这个任务中，我们先通过对谜语的语义、字形、拼音等信息进行分析，从汉字集中召回放入候选池中，之后使用排序模型训练得到谜底的排序，最后使用 top-1 准确率, top-5 准确率, MRR 等作为评判指标。给定的数据包括训练集、验证集和测试集，其中训练集和验证集包含谜语和谜底，测试集只包含谜语，另外还提供了测试集谜底的集合作为参考。

1.2　研究现状

针对解答中文谜语的任务，我们并没有调研到处理完全相同任务的工作，因此我们将范围扩大，调研了部分解决 riddle 问题的任务，这些任务都是处理带选项的解谜问题的，但是其解决问题的思路值得参考。调研的工作总结如下：

1. **Unified QA**[1]. 文章提出的 Unified QA 是一个统一各种问答任务的模型，作者提出将所有类型的数据统一格式，包括不同的谜语、文段、备选项等。在训练过程中，混合选取不同形式的问答任务，目的是保证训练时来自各种类型任务的数据尽可能分布均匀。具体来说，设置一条数据被选中的概率是 $\frac{1}{T_i}$，其中 T_i 代表任务 i 的数据量大小，这样训练过程中可以保证在期望意义下各种任务尽量

均衡。

2. **RiddleSense**[2] 文章介绍了基于英文 riddle 的数据集 RiddleSense，该数据和我们的中文谜语数据集极为相似。文章中提到解决 RiddleSense 问题主要依据 CommonSense 进行推理，属于进阶的自然语言理解任务。为了解决这类问题，文中提出三种常见的方式：fine-tune 已有的 LM 模型；使用符号化的知识图谱进行推理；fine-tune 现有的 T5 进行 text2text 的训练。

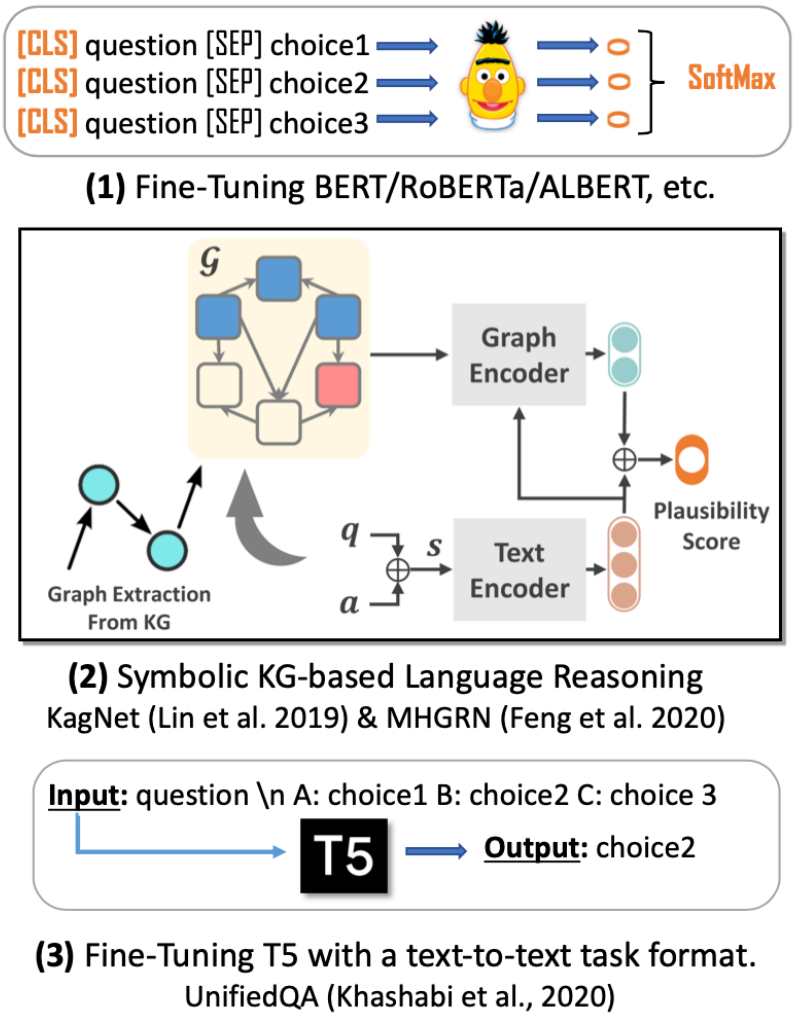

图 1.1　RiddleSense 算法示意图 (来自原论文)

第二章　解决方案

本章介绍中文字谜问题的解决方案，首先通过对样本的分析，找到本问题的难点，提出初步的算法框架，再针对每一项难点给出对应的解决方案。本文最终采用**召回加排序**的算法框架，召回采用**音**、**形**、**义**结合**拆字与拼字**的方法，可以将候选池从整个汉字集缩减到平均几百个，同时覆盖率可以达到 85.2%。排序使用预训练语言模型进行二分类预测的方式实现。

2.1　样本分析

本节介绍样本分析，从训练集中随机抽取 50 个字谜样本，将猜谜方法进行统计分类，结果如图 2.1 所示 (只截取了部分，完整版见附录)。

	拼字法	拆字	同义转换	象形(CV)	语义呼应	谐音呼应	提示字
0 狮　先生来到才鞠躬	1		1	1			
1 烂　秋后又闻王者香	1	1	1				后
2 肴　乘胜向前共一心							
3 嵩　上天峰	1		1				
4 趟　浅水中走过							
5 锻　黄金一段	1						
6 那　一刀子直插乱插							
7 缤　在异乡，为异客	1		1	1			
8 谒　卧龙姓氏却昭著，玄德造庐虔顾三	1		1		1		
9 趵　的的白光照路边	1	1					光、边
10 慎　一直两点，两点一直	1			1			
11 田　雪后初晴							
12 绑　汉高祖还乡容已改	1		1	1			
13 梳　村前流水如书声	1	1				1	前
48 篾　人人方便	1			1			
49 吹　歌后后台说大话					1		
50 伺　横向搞改革							
51 涧　闽江残照飞虹出	1	1			1		
统计	36	12	23	11	5	2	

图 2.1　中文字谜问题训练集部分样本

图中标黄色的谜语是人工难以理解的，共计 10 个，约占 20%，因此本文暂仅考虑其余能理解的 40 个谜语，首先统计是否由**拼字法**求解，所谓拼字法，即根据谜面信息，获取谜面某些字或相关字的偏旁部首，将其从新组合得到谜底的方法。经统计共有 36 个字谜由拼字法求解，占 90% 之多。因此本文认为，中文字谜问题求解的核心方法为拼字法。

拼字法的难点在于，将哪些字的部首拆分，拆分后如何进行组合，组合后如何判断候选谜底的正确性。

2.1.1 拆字法

针对将哪些字的部首拆分，我们进行了详细统计。

首先是谜面中的字，例如谜语“秋后又闻王者香”中，谜底“烂”字的火字旁来源于谜面中的“秋”字，故应将其进行拆字，在 40 个谜语中有 12 个谜语的谜底部首是来源于谜面中的字，占比约 30%，应予以重点考虑。

其次是将谜面进行**同义转换**的字，如上例中的“王者”可以转换成“兰”字，因为有人将兰花誉为“花之王者”，而“兰”是参与谜底拼字的部首。在 40 个谜语中，使用同义转换的谜语有 23 个之多，占比约 58%，也应重点考虑。

由于中文是象形文字，因此谜面的**象形转换**也常被用来提供拆字的元素，例如谜语“先生来到才鞠躬”中“才”的和“扌”为象形关系，再加上“鞠躬”的语义补充，确定了谜底“狮”中“扌”的来源。使用象形方法的字谜共占 11 个，约占 25%。

2.1.2 拆字后的组合

针对拆字后的组合，给定拼字的偏旁部首元素，可以通过查表等方式确定组合后的字。但汉字的组合不一定是唯一的，例如“大”和“丶”既可以组合为“犬”，又能组成“太”。因此对于组合后如何判断正确性就尤为重要。

2.1.3 判断候选谜底的正确性

小小字谜内含了博大的中华文化，再加之其常常需要超常规思维，因此想判断组合成的候选谜底的正确性是较为困难的。本文观察了以下几个角度，给出一些判断依据。

首先是**呼应印证法**，一道好的字谜往往具备巧妙的**语义呼应**或**谐音呼应**。例如字谜“卧龙姓氏却昭著，玄德造庐虔顾三”，谜底为“谒”，使用拼字法，将“卧龙姓氏”同义转换为“诸葛”，进而拆字拼字构成“谒”，由于后半句“玄德造庐虔顾三”有“拜谒”之意，与谜底形成语义呼应，巧妙地印证了“谒”字的正确性。再如字谜“村前流水如书声”，谜底为“梳”，采用拼字法，将“村”和“流”进行拆字拼字构成“梳”，由于后面的“如书声”与谜底形成了谐音呼应，巧妙地印证了“梳”字的正确性。

虽然呼应印证法能够以十分巧妙的方式，精确地印证谜底的正确性，但由于编字谜者的水平不同，能够形成呼应印证的谜题数在数据集中占比很少，经统计，在 40 道谜题中，只有 5 道语义呼应和 2 道谐音呼应的谜题。因此该方法不能作为主要的判断正误的准则。

提示字印证法是另一种判断候选谜底正误的依据，在谜面中很常见，例如“秋后又闻王者香”中的“后”字，就提示了“秋”字要参与拆字，而且要用拆出的后半部分

来拼出谜底。类似的提示字还有很多，诸如“前、后、左、右、边、光、去、离”等，在众多谜语中经常出现，如果模型能学到这些字在拼字法中起的作用，就有可能判断候选谜底的正误。

2.1.4　小结

根据统计，90% 的字谜可以通过拆字加拼字的方法求解，因此本文主要优化拆字加拼字法的性能。由于中文字谜通常考虑“音形义”，在拆字阶段要分别设计对应模块处理。拆字后的组合不是难点，而判断候选谜底的正确性是中文字谜问题最大的难点。其中呼应印证法准确性最高，但适用范围过小，本文暂未考虑使用，而提示字法在谜语中的出现频率较高，对谜底的印证能力也较好，可以使用模型进行学习。

2.2　中文字谜求解框架

本文提出的中文字谜求解框架采用**召回加排序**的方式。其中**召回**主要解决上节提到的**将哪些字的部首拆分**和**拆分后如何进行组合**两个问题，**排序**解决**判断候选谜底的正确性**的问题。

2.2.1　召回

考虑音形义的拆字算法

为了尽可能增加谜底召回率，对谜语的义、形、音都进行了分析后，我们发现这三种信息对于解谜都是有帮助的，于是在此基础上尝试了不同的召回方法。例如，对于谜语“先生来到才鞠躬”，我们需要根据“先生”的语义联想到“师”，根据“才鞠躬”的形联想到“犭”，同时利用字形字义，合起来才能得到谜底“狮”。对于谜语“古稀回到故园内”需要使用“园”字的音“*yuan*”，才能联想到“袁”，继而得到谜底“辕”。

对于字义的利用，我们使用了 **Chinese Semantic KB**[3]的同义词词典作为知识来源，对每一个字谜利用 **jieba** 进行分词后，再查找每一个词语的近义词，最后将这些近义词中出现频率最高的十个字作为拆字的候选。

对于字形，我们计算得到与字谜中每个字字形最近似的十个字或偏旁部首作为拆字的候选。对于两个字之间的字形相似度，我们借鉴了 **nlp-hanzi-similar**[4]的方法，具体的计算过程是先笔画数相似度、结构数相似度、四角编码相似度和偏旁部首相似度，最后通过自定义的权重加权计算得到一个分数。其中，形近字字典的来源包括训练集、验证集和测试集所有字及其拆字结果。

对于字音，我们使用了 **chinese homophone char 词典**[5]，对于字谜中的每一个字，先利用 **pypinyin** 这一工具包得到拼音，再根据拼音得到每个字出现频率最高的十个同

音字（若没有十个以上同音字，则把全部同音字加入拆字候选）加入拆字候选。

拼字算法与候选生成

分析谜语的音形义后，我们将谜面的汉字以及谜语的同音字、形近字、同义字进行拆字拼字得到谜底。根据我们对样本的分析，有些样本单单利用谜面拆字拼字就能得到谜底，但大部分谜底需要同时利用音形义转换中的多个组合才能求解，因此我们选取不同的集合进行拆字，从而得到谜语的候选集。

我们对比了四种组合：纯谜面、加同义转换 (yi)、加象形和同义转换 (xingyi)、加同音同义和象形转换 (yinxingyi)。具体的拼字算法如下：

- **Step 1**. 对某个谜语考虑音形义得到的所有汉字进行拆字（考虑四种组合）得到部首集 a。
- **Step 2**. 遍历谜底集进行召回，即将谜底集的每个汉字进行拆字得到部首集 b，在 Step 1 中得到的部首集 a 中查找 b 在 a 中的比例，比例大于某个阈值就将该汉字添加进谜底的候选池中。

这里比例的阈值是人为设置的，阈值越低，候选池中的汉字越多，一般这种情形下覆盖率更高，但是对排序模型的压力也会升高。

2.2.2 排序

本文将排序任务视为二分类任务，即训练一个二分类器，输入为谜面和一个候选谜底，输出为该候选谜底正确的概率。

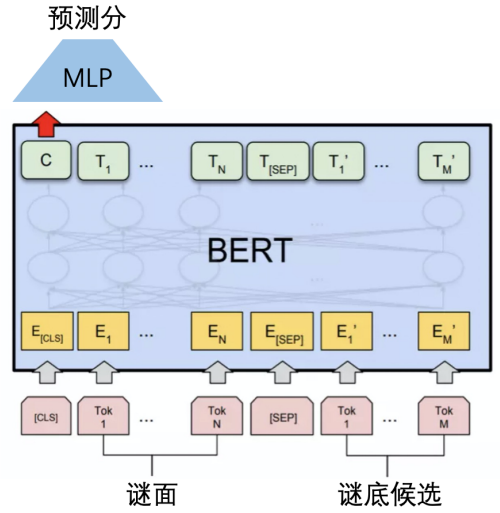

图 2.2 排序模型结构

排序模型选用预训练语言模型 BERT，由于 BERT 的预训练任务中有一项为预测两个句子间关系，因此为了获得较好的 finetune 效率，本文采用类似的输入方式，将谜

面和谜底当做前后两个句子，中间用 [SEP] 分隔符分割，通过 BERT 进行编码。之后使用 [CLS] 向量作为特征向量输入到一个 MLP 分类器中。最后使用 sigmoid 计算概率，使用交叉熵损失进行模型更新。整体模型结构如图 2.2 所示。

对谜面和谜底候选的输入方式，本文进行了若干尝试：

- 1. 谜面用原文，谜底用原文
- 2. 谜面用原文，谜底用其拆成部首的序列
- 3. 谜面用原文 + 其拆成部首的序列，谜底用其拆成部首的序列

实验发现，采用第 2 种方式的效果最优，细节及分析将在下一章介绍。

第三章　实验结果与分析

3.1　实验设定

数据集分为 train、valid 和 test，其中 train 和 valid 带有谜底标签，test 无标签，但提供了谜底集，由于 valid 集较大，为了提升训练效率，将 valid 集随机采样至 2000 条样本，构成 valid2000。为了测试模型在 test 集上的性能，我们从 test 集上随机抽取 100 个样本进行人工标注，构成数据集 test100。数据构成如表 3.1 所示。

表 3.1　中文字谜问题数据集构成

数据集	样本数	标签集大小
train	16631	4371
valid	5480	1458
valid2000	2000	974
test	5413	1458
test100	100	1458

实验测试了有召回的方法和无召回的方法，对于有召回的方法，候选集为召回提供，无召回的方法，候选集为全体标签集。训练的正样本取训练集的谜面和标签，负样本按照 1:9 的正负样本比进行负采样，其中无召回的方法在 train 标签集进行随机负采样，有召回的方法在候选集进行随机负采样。

对于无召回的方法，针对是否对谜面和谜底拆部首进行了实验，并尝试了不同的预训练语言模型。对于有召回的方法，对比了考虑不同召回因素组合的优劣。

实验的评价指标选择 top1 准确率、top5 准确率和 top5 的 MRR 指标。

3.2　实验结果及分析

中文字谜问题的实验结果如表 3.2 所示。

表 3.2　中文字谜问你题实验结果

方法	valid2000			test100			
	acc@1	acc@5	MRR@5	acc@1	acc@5	MRR@5	rec fail
BERT-radicle	0.1	0.6	0.27	0.0	0.0	0.0	-
BERT-noradicle	0.1	0.4	0.25	0.0	0.0	0.0	-
BERT-base	18.9	**32.2**	**24.1**	12.0	**27.0**	17.6	-
Ernie-base	**19.3**	31.4	23.9	11.0	19.0	14.5	-
BERT-rec	10.1	18.2	13.3	6.0	16.0	9.5	45.1
BERT-rec-yi	12.2	21.7	15.8	8.0	13.0	10.1	37.9
BERT-rec-xingyi	13.1	23.7	17.0	10.0	21.0	13.8	25.1
BERT-rec-yinxingyi	17.1	28.2	21.4	**15.0**	22.0	**17.9**	14.8

表中上半部分为无召回的方法，方法 BERT-radicle、BERT-noradicle 和 BERT-base 分别对应谜面和谜底都拆成部首序列、都使用原文、谜面原文谜底拆部首的三种做法。可以看出前两种做法的效果非常差，而最后一种效果很好。这是由于排序模型需要从谜面中得到语义信息，因此谜面要使用原文；对于谜底，由于 train、valid、test 的谜底集无交集，直接使用谜底原文进行训练缺失了泛化性，而将谜底拆成部首后，既给模型提供了拆字拼字法的要素，又在 train、valid、test 的谜底集间搭起了桥梁，增加了泛化性，因此获得了较好的性能。对于不同的预训练语言模型，本文尝试了 BERT 和 Ernie，实验结果表明二者的效果相差不大。

表中下半部分为有召回的方法，其中 BERT-rec 仅使用谜面中的文字进行拆字并组合生成候选集，召回失败率较高，为 45.1%。而 BERT-rec-yi、BERT-rec-xingyi、BERT-rec-yinxingyi 三个方法的在 BERT-rec 的基础上分别增加了“义召回”、“形义召回”、“音形义召回”，召回失败率也逐渐降低。分别使用不同的召回集进行训练并测试得到表中结果。可以看出随着考虑召回因素的增加，召回失败率在降低，模型的效果也在提升。

对比有召回和无召回的方法，有召回的方法在 valid2000 上的表现明显差于无召回的方法，在 test100 上也只是略高一点，推测其原因本文认为有两点。

第一，排序模型本身过于强大，而且泛化性很强。而召回模型是使用规则做的，缺乏泛化性。召回模型主要考虑的是中文的“音形义”加拆字拼字，对于预训练语言模型，已经通过大规模语料学到了“义”，而根据第二章的统计，“音形”的占比远没有“义”高。同时给模型输入谜底的部首序列又帮助模型学习了拆字拼字，所以排序模型也可以在一定程度上直接学到中文字谜求解的方法，再加之其强大的泛化性，就显得

召回模型不那么重要了。

第二，本任务的 test 集提供了谜底集，而现实环境下的猜谜任务是不会提供谜底集的，因此如果不进行召回的话，排序算法必须在 20000 多个汉字上进行排序，这个难度就远非在目前 1000+ 的谜底集上做排序可比的了。

所以本文认为，对于中文字谜问题，召回模型有其存在的重要性和必要性，对于已给较小谜底集的场景，由于排序模型的强大，召回模型的重要性被隐去了，但在真实场景中，由于没有谜底集的限制，排序模型必须依靠召回模型缩小候选范围，继而更好地发挥排序的作用。

参考文献

[1] KHASHABI D, MIN S, KHOT T, et al. Unifiedqa: Crossing format boundaries with a single qa system[J]. ArXiv preprint arXiv:2005.00700, 2020.

[2] LIN B Y, WU Z, YANG Y, et al. RiddleSense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge[J]. ArXiv preprint arXiv:2101.00376, 2021.

[3] Liuhuanyong. ChineseSemanticKB[EB/OL]. https://github.com/liuhuanyong/ChineseSemanticKB.

[4] Houbb. Nlp-hanzi-similar[EB/OL]. https://github.com/houbb/nlp-hanzi-similar.

[5] LI L. ChineseHomophones[EB/OL]. https://github.com/LiangsLi/ChineseHomophones.

附录 A　附录

		拼字法	拆字	同义转换	象形(CV)	语义呼应	谐音呼应	提示字	
0	狮 先生来到才鞠躬	1		1	1				
1	烂 秋后又闻王者香	1	1	1				后	
2	肴 乘胜向前共一心								
3	嵩 上天峰	1		1					
4	趟 浅水中走过								
5	锻 黄金一段	1							
6	那 一刀子直插乱插								
7	缤 在异乡，为异客	1		1	1				
8	谒 卧龙姓氏却昭著，玄德造庐虔顾三	1		1		1			
9	趵 的的白光照路边	1	1					光、边	
10	慎 一直两点，两点一直	1			1				
11	田 雪后初晴								
12	绑 汉高祖还乡容已改	1		1	1				
13	梳 村前流水如书声	1	1				1	前	
14	误 子游对曰	1		1					国籍
15	申 神州要统一								
16	机 春色清风里	1	1	1				清_里	
17	赫 近朱者赤	1		1					
18	女 人说多子为好，我说少生为妙	1							
19	辍 双双重逢近古稀	1		1					老爷车
20	荐 含羞夜半会一人			1					
21	焐 欲语无言听秋声		1					无x	
22	社 埋没里边，视而不见	1						没x，不x	
23	杌 独子植树	1		1					
24	梦 桑榆暮景忆邯郸	1		1					暮-夕阳
25	公 山痕宛宛能助眉丰								
26	萌 一朝改革立见生机		1			1		改革	
27	辕 古稀回到故园内	1		1			1		
28	黧 利字当头不算黑	1							
29	翠 八千子弟兵								
30	鹐 鹦鹉嘴上衔弓箭	1		1	1				
31	羊 一叠人民币				1				
32	舄 臼中有鸟飞不还	1							
33	颍 残花顺水过前川	1	1					残	
34	栌 卢前残月照楼头	1	1					头	
35	静 雨后春笋								
36	灾 容貌欠端方	1			1			欠	
37	隋 孤帆斜风里，江水落玉盘	1	1	1	1			里、落	
38	跌 人生没有单行道	1		1					
39	互 上下别扭				1				
40	倚 香凝赤心留人间								
41	苣 开渠送水放草下	1		1				送	
42	炳 灾后两人杳无踪	1	1					后 杳无踪	
43	格 春归在客先	1	1	1				先	
44	伤 一人无力旁边站，一人有力上边躺，此事	1		1	1	1			卧人
45	鲜 两个动物并排站，一个游泳，一个吃草。	1		1					
46	悻 幸福在于心相随	1		1					心
47	樗 下雪之后双梅开	1		1					雪
48	篾 人人方便	1			1				人 竹
49	吹 歌后后台说大话					1			
50	伺 横向搞改革								
51	涧 闽江残照飞虹出	1	1			1			
54	统计	36	12	23	11	5	2		

图 A.1　中文字谜问题训练集样本分析