

## Introduction

In this project, a popular task in many primary school Chinese exams, disarranged sentence reconstruction (词句重组), is studied. This task mainly requests the students to reorder and re-arrange a list of Chinese words in random order to form a grammatically correct and meaningful sentence.

村庄 浓密 的小湖 树林里 那边的 掩隐 在

小湖那边的村庄掩隐在浓密的树林里

The project's aim is to automatically solve such problems. To fit the amount of work within the constraint of the project, we currently limit our vocabulary to those words which commonly occur in primary school texts.

The final deliverable of this project is a Python program that run at the back-end of a website. You can input to the program disarranged word lists through the webpage, and view the answers generated by the program. Scan the QRCode on the right to access the website, or directly go to the address:

<http://tinyurl.com/centaddsr>



Our project attempts to understand the word order conventions in natural languages. Such understanding can be of crucial importance to various natural language processing applications. For example, this algorithm, if fully trained with a larger scope of text data, may be useful as an add-on that can increase the performance of our current machine translation systems by correcting the words order in the translation result.



程序可以自动解决此类问题具有较高的精度。

Expected: 程序可以自动以较高的精度解决此类问题。

## Models

### Word-based N-gram language model

$P((i)^{\text{th}} \text{ word} = \text{"xxx"} | (i-n \sim i-1)^{\text{th}} \text{ words} = \text{"xxx xxx xxx"})$

我正在看小说  
他最喜欢的是看书  
我们要多看书才能长知识

$P((i)^{\text{th}} = \text{"书"} | (i-1)^{\text{th}} = \text{"看"}) = 2/3$

### Word-based Backward N-gram language model

$P((i)^{\text{th}} \text{ word} = \text{"xxx"} | (i+1 \sim i+n)^{\text{th}} \text{ words} = \text{"xxx xxx xxx"})$

这真是一朵漂亮的花  
那本书有着漂亮的封面  
她穿得真漂亮呢

$P((i)^{\text{th}} = \text{"漂亮"} | (i+1)^{\text{th}} = \text{"的"}) = 2/3$

### Char-based N-gram language model

Use characters as basic units to consider the similarities between different words, hence account for rare words

他高兴地跳了起来  
小明高兴地笑了

大家高高兴兴地去上学

High Possibility

High Possibility

### POS-based N-gram language model

Use part-of-speech instead of words to consider similarities between different words, hence account for rare words

(Assume the program does not know the word 阅读)

我喜欢阅读故事

???

n. + v. + v. + n.

High Possibility

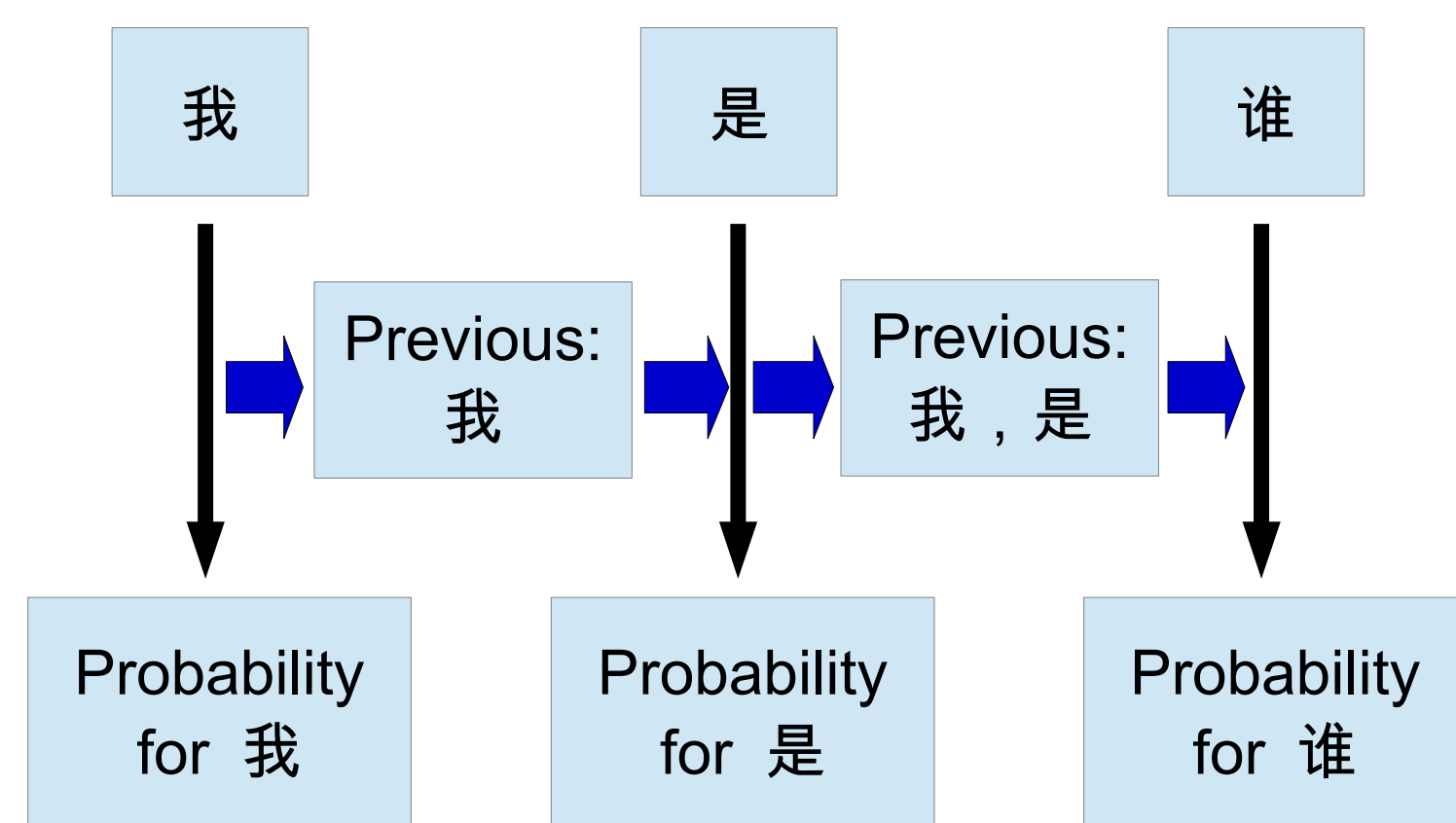
### Bigram Occurrence language model

Emphasis on common collocations like “数学书”

$n(\text{bigram})$  = occurrence of *bigram* in data  
 $\text{score}(\text{bigram}) = [\log n(\text{bigram})]^2 + 1$

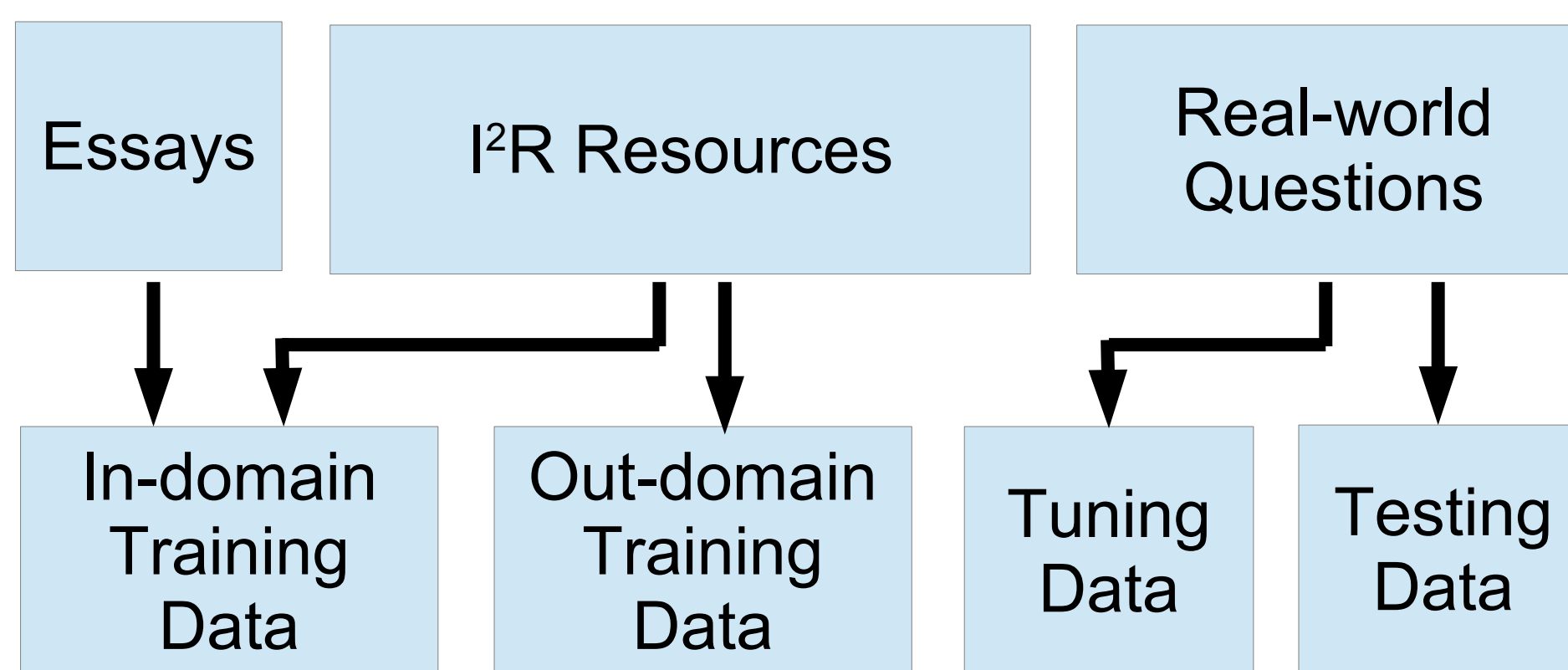
### Recurrent Neural Network language model (RNN)

Evaluate each word based on the full history



## Data Collection

Large amount of Chinese text for the training and verification of statistical models



## Programming

In order to determine the best hypothetical answer to a task, we maximise a score, which is a weighted summation of scores in different models

$$\text{score}(\text{hypothesis}) = w_1 s_1 + w_2 s_2 + w_3 s_3 + \dots + w_n s_n$$

The maximum is approximated by adding words to the hypothesis one by one, using the following pseudocode:

```
function generate_hypothesis:
  init hypos to empty array
  hypos.append empty string
  for each index in 1 to n:
    init new_hypos to empty array
    for each hypothesis in hypos:
      for each word in hypothesis.remaining_words:
        new_hypos.append (new hypothesis)
      end
    end
    new_hypos.sort by score
    if new_hypos.contains more than 100 items:
      set hypos to first 100 items in new_hypos
    else:
      set hypos to all items in new_hypos
    end
  end
  return first item in hypos
end
```

### Generation Directions:

Left-to-right

Old hypothesis + Word = New hypothesis

Right-to-Left

Word + Old hypothesis = New hypothesis

Bi-directional

Old hypothesis + Word = New hypothesis 1

Word + Old hypothesis = New hypothesis 2

### Chinese Primary School Essays:

- ~11M characters original, filtered to 340k sentences.
- Crawled from “作文网” (www.zuowen.com) with Python script

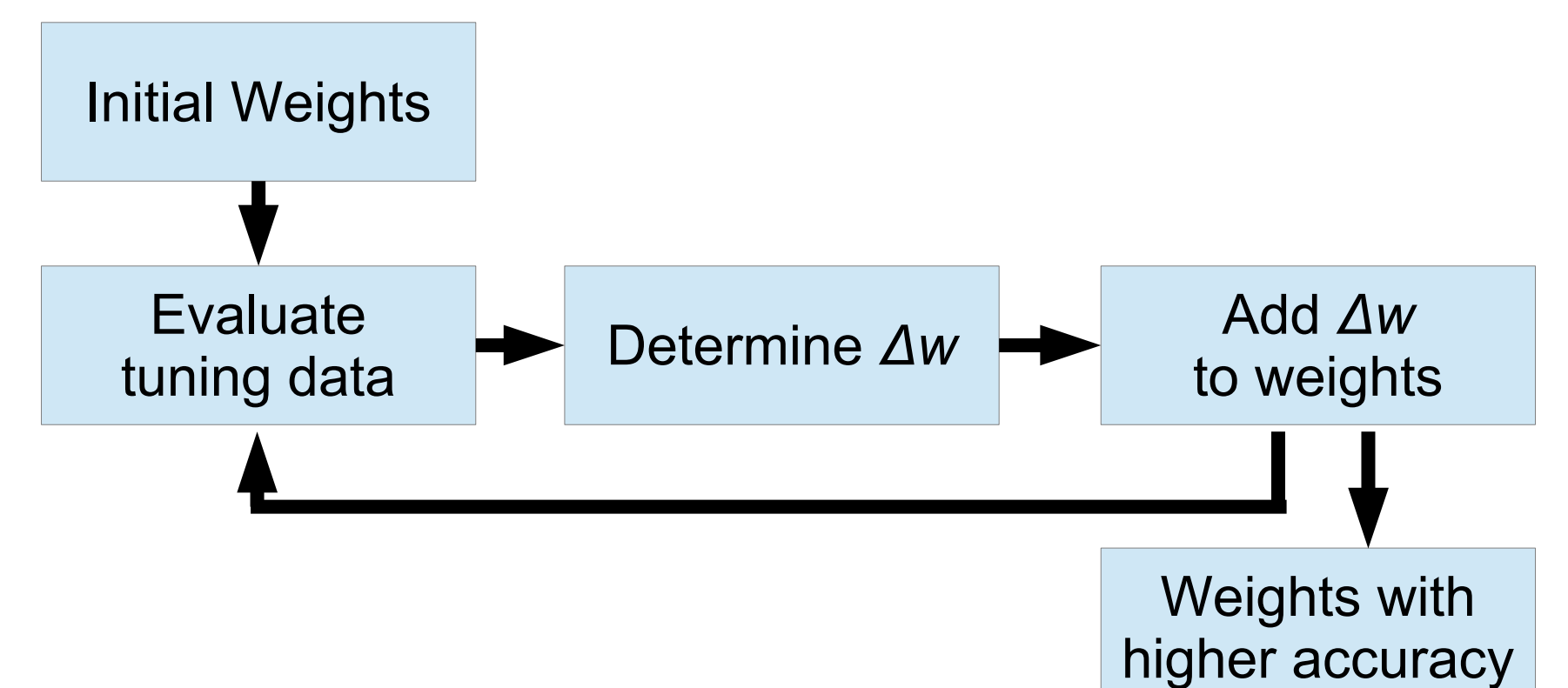
### I2R Resource

- 15M sentences that are partially relevant

### Real-world 词句重组 questions and answers

- 2033 sentences, split into 1500 lines of tuning data and 533 lines of testing data
- Collected from Chinese education websites and digital versions of exam papers online

## Tuning



**Evaluation of hypothesis:** BLEU (BiLingual Evaluation Understudy) is used to compare generated hypotheses to correct answers. It considers the precision of the hypothesis, or the percentage of phrases of a certain length in the hypothetical sentence that also appear in the correct answer.

Machine Output: 我喜欢读这本书漫画

Correct answer: 我喜欢读这本漫画书

Phrase Length	1	2	3	4
Total in hypothesis	6	5	4	3
In correct answer	6	3	2	1
Precision	1	0.6	0.6	0.3

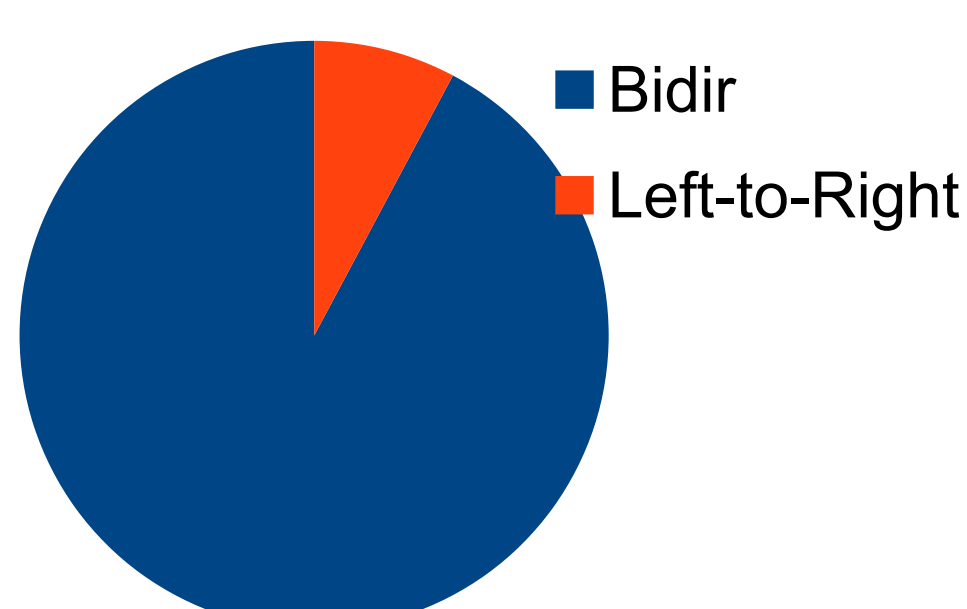
**Tuning methods:** efficient means of determining  $\Delta w$  that can result in an improvement in accuracy

- Minimum Error Rate Tuning (MERT): Minimise BLEU difference between best hypothesis and correct answer
- Margin Infused Relaxed Algorithm (MIRA): Maximise score difference between hypothesis closest to correct answer and other hypotheses
- Alternative Tuning: switch between the two algorithms iteration by iteration to benefit from both algorithms

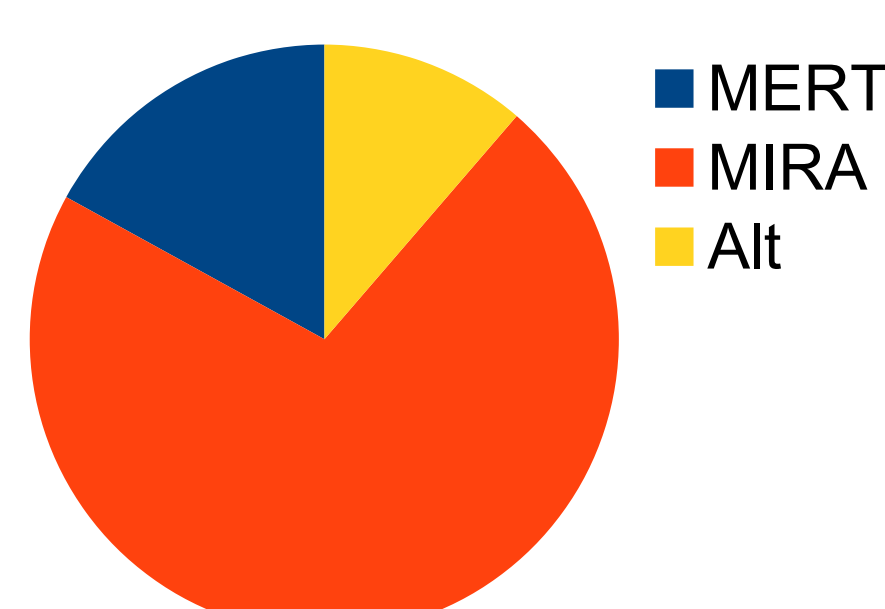
## Results

### Hypothesis Generation Direction

### Tuning methods



Right-to-Left does not show very good results, due to the fact that most languages form their sentences from left to right



Adjusting tuning methods causes improvements as small as 0.005, hence such adjustment is of the least importance

### Comparison between models

Evaluating RNN language model is quite time-consuming, but is able to introduce the greatest improvement in accuracy. Hence, RNN should be used when computational requirements can be met.

Most other language models are useful as well, as there is a general trend that the more models we include, the better the results we can get.

Some not-so-useful models:

- Backward n-gram language model: due to words are still more related to their precedents
- Models built from lexicons: high irrelevance to primary school context

## Conclusion

The project is limited by the amount of data collected, which may lead to overfitting. However, it still succeeded in solving the famous task in many primary school Chinese exams, namely disarranged sentence reconstruction, automatically, and all the scripts in this project can be easily re-run on a larger data set to obtain more accurate results.

In the future, the project can be extended to larger scopes and even other languages, if enough data is made available to the program. We believe that the findings in this project can be very useful in such extensions.