浅谈科研训练对项目实践的指导作用

以"现代诗歌生成"问题为例

报告人：陈峻
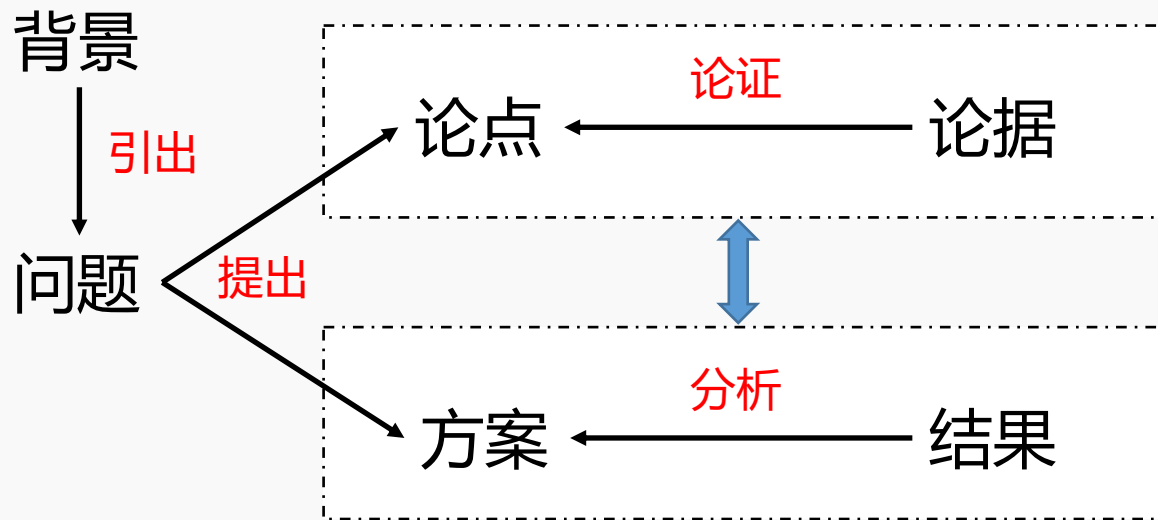
介绍：http://iir.ruc.edu.cn/~chenj/index.html

时间：2018.11.08

Overview

High

The level of difficult

Low

Boundaries — The ability to know the boundaries of the problems and methods

Solutions — The ability to solve the problems

Methodologies — The ability to master the methodologies

The ability of research
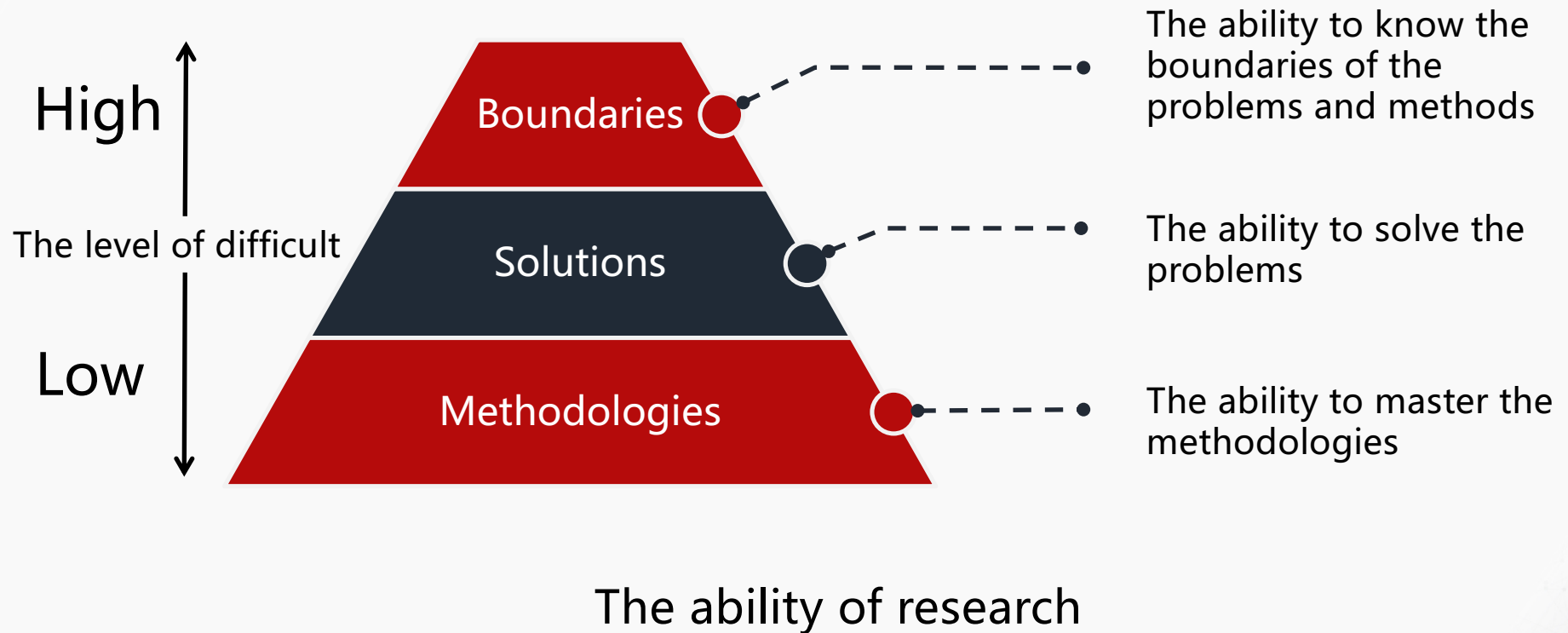
Methodology

打辩论　　写材料　　写论文　　解决问题　　...

背景

引出

问题

提出

论点 ← 论证 — 论据

方案 ← 分析 — 结果

背景："魏则西事件"等

辩题：搜索引擎中的竞价排名技术该不该被禁止？

总论点：搜索引擎中的竞价排名技术竞价排名不该被禁该。

分论点：竞价排名技术本身没有错，错在缺失相关法律监督。

论据1：竞价排名被广泛地应用在其他相关行业，这些行业行业有明确的法律规定，如拍卖。若没有明确的法律条文，仅通过道德层面去约束一项技术，或者一刀切地禁止这种技术是极其欠妥的。比如，举例说明...

Methodology

介绍背景（Background）

背景：搜索引擎搜集的页面很多，要如何返回满足用户需求的页面。

为什么（why）

引出问题（Problem）

问题：搜索引擎中的页面排序问题
输入：关键词
输出：相关的页面

是什么（what）

提出方案（Solution） ← 分析（Analyze） 获取结果（Result）

理论（Theoretical）

要怎样（how）

实验（Experimental）

会怎样（how）

提出方案：Pagerank算法

分析结果：用户满意度标注

总结（Discussion） 展望（Future work）

Methodology

Real scenarios

Others

Not addressed before

Addressed before

Research problem

Has challenges

Has no challenges

Novel methods

Not novel methods

Agenda of an instance
Modern poetry generation

- Background
- Problem
- Solution
  - State-of-the-art methods
  - Our method and implementation
- Result
  - Experimental setups
  - Experimental studies
- Discussions and future works
- References

Background

- [Generating Chinese Couplets using a Statistical MT Approach, Jiang et al., Coling 2008]
  - Part of the unique cultural heritage of China is the game of Chinese couplets (duìlián). One person challenges the other person with a sentence (first sentence). The other person then replies with a sentence (second sentence) equal in length and word segmentation, in a way that corresponding words in the two sentences match each other by obeying certain constraints on semantic, syntactic, and lexical relatedness. This task is viewed as a difficult problem in AI and has not been explored in the research community.

- [Chinese Poetry Generation with Planning based Neural Network, Wang et al., Coling 2016]
  - The classical Chinese poetry is a great and important heritage of Chinese culture. During the history of more than two thousand years, millions of beautiful poems are written to praise heroic characters, beautiful scenery, love, friendship, etc. There are different kinds of Chinese classical poetry, such as Tang poetry and Song iambics. Each type of poetry has to follow some specific structural, rhythmical and tonal patterns. Table 1 shows an example of quatrain which was one of the most popular genres of poetry in China. The principles of a quatrain include: The poem consists of four lines and each line has five or seven characters; every character has a particular tone, Ping (the level tone) or Ze (the downward tone); the last character of the second and last line in a quatrain must belong to the same rhyme category (Wang, 2002). With such strict restrictions, the well-written quatrain is full of rhythmic beauty. … As a consequence, the existing methods may fail to generate meaningful poems if a user wants to write a poem for a modern term (e.g., Barack Obama).

Problem

- Input
  - A sequence of keywords
    - e.g., 相处, 永久, 并肩, 朦胧, 思念, 真心真意
- Output
  - A modern poetry matching the topics described by the keywords
    - e.g., 两个人相处，才能到永久。与你肩并肩，渐渐朦胧。深深的思念，我对你真心真意。
- Requirement (challenge)
  - meaningful
  - relevant
  - emotional
  - consistent
  - creative

无题
(JD)

两个人相处 (相处)
才能到永久 (永久)
与你肩并肩 (并肩)
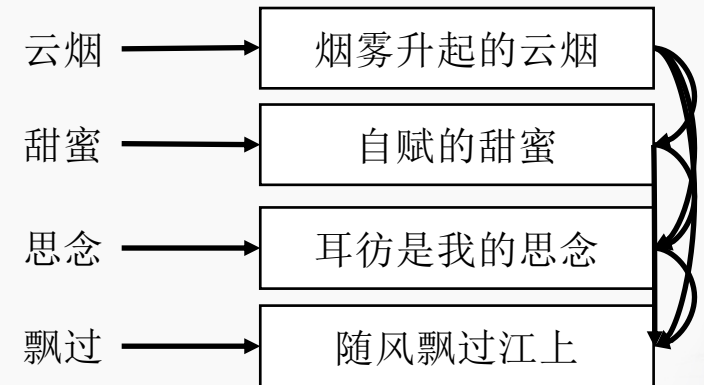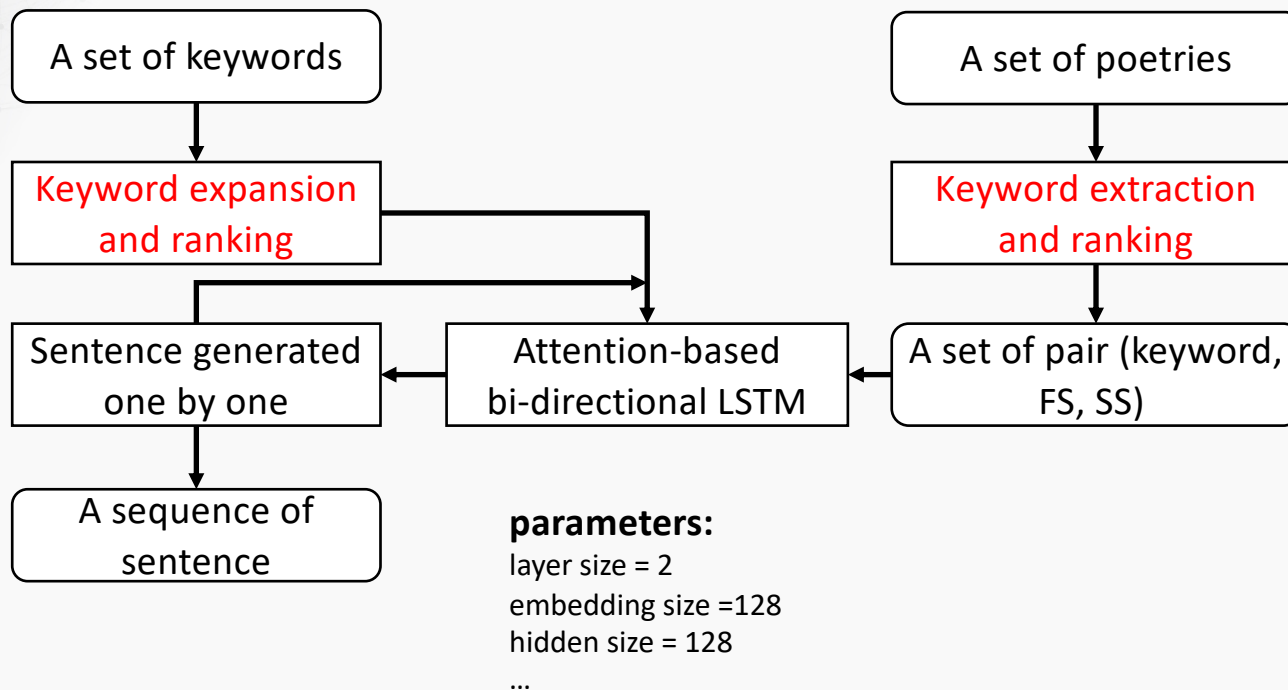渐渐朦胧 (朦胧)
深深的思念 (思念)
我对你真心真意 (真心真意)

Solution
State-of-the-art methods

- Retrieval-based methods
  - phrase search (Tosa et al., 2009; Wu et al., 2009)
  - word association norm (Netzer et al., 2009)
  - genetic search (Zhou et al., 2010)
  - template search (Oliveira, 2012)
  - text summarization (Yan et al., 2013)

- Generation-based methods
  - statistical machine translation (Jiang and Zhou, 2008; He et al., 2012).
  - sequence-based RNN (Zhang et al., 2014)
  - character-based RNN (Wang et al., 2016)
  - Bi-directional LSTM, GRU (Wang et al., 2016)
  - Attention mechanism (Wang et al., 2016)

Solution
Our methods and implementation

A set of keywords

Keyword expansion and ranking

Sentence generated one by one

A set of poetries

Keyword extraction and ranking

Attention-based bi-directional LSTM

A set of pair (keyword, FS, SS)

A sequence of sentence

parameters:
layer size = 2
embedding size =128
hidden size = 128
…

云烟 → 烟雾升起的云烟

甜蜜 → 自赋的甜蜜

思念 → 耳彷是我的思念

飘过 → 随风飘过江上

Generation-based framework
https://github.com/Disiok/poetry-seq2seq

Solution
Our methods and implementation

A set of keywords

↓

Keyword expansion and ranking

↓

word2vector → Sentence ranking
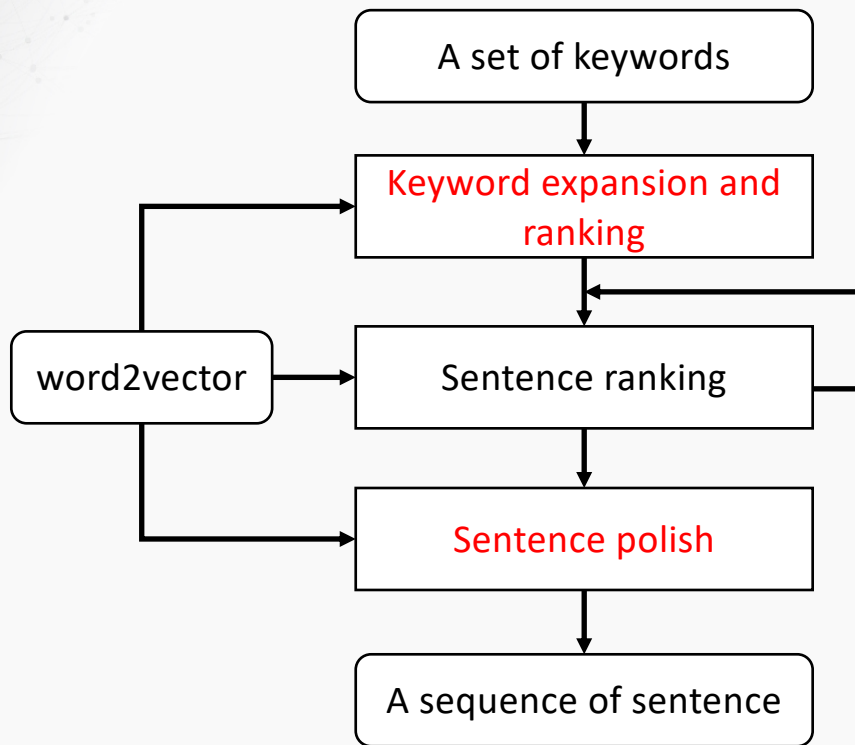
$$R(S_i|Q_i, S_{i-1}, \ldots, S_{i-l}) = \lambda \sum_{m=0}^{|S_i|} \sum_{n=0}^{|Q_i|} r(S_i^m, Q_i^n) + (1-\lambda) \sum_{j=i-1}^{j=i-k} \frac{1}{e^l} \sum_{m=0}^{|S_i|} \sum_{n=0}^{|S_j|} r(S_i^m, S_j^m)$$

$$r(x, y) = idf(x) * sim(x, y)$$

↓

Sentence polish

↓

A sequence of sentence

Retrieval-based framework

Result
Experimental setups

- Datasets
  - training data
    - 9k poetries + 4.5k poetries
  - test data
    - 20 groups of keyword sequence
      - keyword:
        - 相处, 永久, 并肩, 朦胧, 思念, 真心真意
      - poetry:
        - 两个人相处，才能到永久。与你肩并肩，渐渐朦胧。深深的思念，我对你真心真意。
- Metrics
  - subjective evaluation
    - 4 human beings (2 JD NLP experts, 2 AI Camp students) manually judge the syntactics, semantics, creativities

Result
Experimental studies

	Team 1	Team 2	Team 3
Sum (300)	**211**	200.25	202.25
Avg (15)	**10.55**	10.01	10.11
#poetries > 8 (20)	16.5	**17.5**	16.5

Result
Experimental studies

- Use Cases

Case 1
(Generation-based model)

浓密的树林 (树林)
松疏的绿色 (桔红色)
开始公交车 (旅途)
补帛宁静的波动 (宁静)
低矮的屋檐 (屋檐)
玲珑迷迷人 (迷人)
静港之乡 (思乡)
又是夕夕的神话 (神话)
整整夜晚 (夜晚)
妒眼汪汪悴汪 (眼泪)

Case 1
(Retrieval-based model)

树林的幽语 (树林)
音乐把宁静铺在桔红色的心情上 (桔红色)
梦枕着旅途的疲惫 (旅途)
夜开始宁静 (宁静)
滴嗒的雨声敲在故乡的屋檐 (屋檐)
绕过的湖滨依然散落着相思的迷人 (迷人)
思乡的故事永远的那麽流畅 (思乡)
象一个神话对孩子的永恒 (神话)
那一个夜晚 (夜晚)
它是我的眼泪 (眼泪)

Result
Experimental studies

- Use Cases

<div style="display:flex">

Case 3
(Generation-based model)

记忆都是愚蠢的记忆 (记忆)
深沉的夜晚 (夜晚)
潺潺的歌声 (声音)
这片土地呼啸的土地 (碎片)
钻进我们的庭院里 (扔进)
强光 (强光)
黄土泥土被谁打丝锈的 (铁片)

Case 3
(Retrieval-based model)

记忆不再等我 (记忆)
很多夜晚 (夜晚)
我是我的土地的声音 (声音)
即使是碎片 (碎片)
并扔进那将其 (扔进)
融为一体的白昼的强光中时 (强光)
象响亮的烂铁片 (铁片)

</div>

Result
Experimental studies

- Use Cases

Case 17
(Generation-based model)

烟雾升起的云烟 (云烟)
自赋的甜蜜 (甜蜜）
耳彷是我的思念 (思念）
随风飘过江上 (飘过)
遗失了过错了的 (消失)
烟烟云烟 (云烟)
扶动我翼翼的心弦 (心弦)

Case 17
(Retrieval-based model)

你是四月早天里的云烟 (云烟)
是这样的甜蜜 (甜蜜)
我把您思念 (思念)
她飘过 (飘过)
如果在它们消失的时候 (消失)
你是四月早天里的云烟 (云烟)
心弦呵 (心弦)

Discussions and future works

- Bridge the retrieval-based methods and generation-based methods
  - improve keyword-to-topic expansion beyond the tedious combination of keywords
    - considering the roles of each keyword in poetry generation
  - improve keyword-to-sentence matches beyond only applying a similarity measure (meaningful, relevant, emotional)
    - combining diverse similarity measures
    - mixing the NN-based methods
  - improve sentence-to-sentence coherence and fluency beyond separately matching to the keyword (consistent)
    - considering more historical context cues
    - considering the key topic of a poetry
  - improve creativity of sentence beyond only matching the existing corpus (creative)
    - generating similar sentence from the retrieved sentences
    - replacing some keywords

Conclusions

- Useful tips for taking away
  - methodologies to supervise your researches
    - practice to conclude what you read
  - skills to write/express your ideas
    - practice to write/express your ideas from "small" to "large"
    - writing in Chinese vs. writing in English
  - strategies to start your problems
    - related work survey first vs. think by yourself first
    - communicate with your advisers
  - ways to improve your abilities for solving problems
    - implement the solutions and observe the empirical results
    - think about the strengths and drawbacks
    - learn to obtain the intuitions and insights
  - …

References

- Xingxing Zhang, Mirella Lapata: Chinese Poetry Generation with Recurrent Neural Networks. EMNLP 2014: 670-680

- Qixin Wang, Tianyi Luo, Dong Wang, Chao Xing: Chinese Song Iambics Generation with Neural Attention-Based Model. IJCAI 2016: 2943-2949

- Jiyuan Zhang, Yang Feng, Dong Wang, Yang Wang, Andrew Abel, Shiyue Zhang, Andi Zhang: Flexible and Creative Chinese Poetry Generation Using Neural Memory. ACL (1) 2017: 1364-1373

- Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, Enhong Chen: Chinese Poetry Generation with Planning based Neural Network. COLING 2016: 1051-1060

- Rui Yan, Han Jiang, Mirella Lapata, Shou-De Lin, Xueqiang Lv, Xiaoming Li: i, Poet: Automatic Chinese Poetry Composition through a Generative Summarization Framework under Constrained Optimization. IJCAI2013: 2197-2203

- Hugo Gonçalo Oliveira: A Survey on Intelligent Poetry Generation: Languages, Features, Techniques, Reutilisation and Evaluation. INLG 2017: 11-20

- Long Jiang, Ming Zhou: Generating Chinese Couplets using a Statistical MT Approach. COLING 2008: 377-384

- Naoko Tosa, Hideto Obara, Michihiko Minoh: Hitch Haiku: An Interactive Supporting System for Composing Haiku Poem. ICEC 2008: 209-216

- Yael Netzer, David Gabay, Yoav Goldberg, and Michael Elhadad. 2009. Gaiku: Generating haiku with word associations norms. In Proceedings of the Workshop on Computational Approaches to Linguistic Creativity. Association for Computational Linguistics, pages 32–39.

- Jing He, Ming Zhou, Long Jiang: Generating Chinese Classical Poems with Statistical Machine Translation Models. AAAI 2012

Thanks!
Q&A