iFLYTEK 多语种文本挖掘挑战赛

团队：肉蛋葱鸡

赵嘉豪[1]，赵达达[2]

[1]中科院自动化所

[2]西安电子科技大学

zhaojiahao2019@ia.ac.cn

ddzhao@stu.xidian.edu.cn

2020年10月

大纲

- 团队
- 整体方案
- 后续优化
- 总结

大纲

- 团队
- 整体方案
- 后续优化
- 总结

团队简介

- 队伍名：肉蛋葱鸡
  - 梗
  - 肉蛋冲击 -> 肉蛋葱鸡
  - 网络环境下语言的变化性
- 成员
  - 赵嘉豪，中科院自动化所，直博二年级
  - 赵达达，西安电子科技大学，硕士二年级

游戏主播：芜湖大司马[1]

[1] https://h.bilibili.com/93000131

大纲

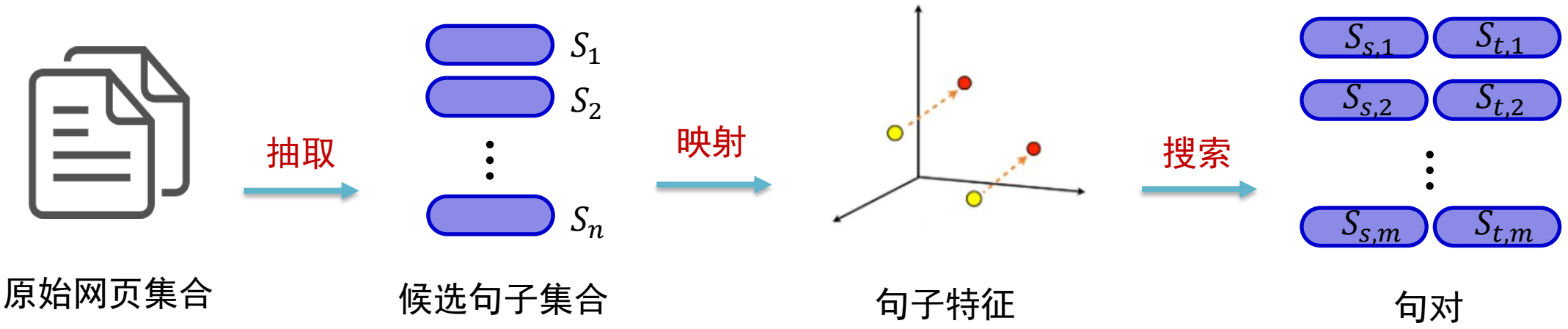
- 团队
- 整体方案
- 后续优化
- 总结

赛题任务

- 从多个语种的单语网页原始语料，抽取双语平行句对
  - 任务
    - 初赛：（中文，日语）　　复赛：（中文，意大利语）
  - 数据
    - 初赛：86GB　　复赛：20GB
  - 要求
    - 禁止使用任何其他数据，禁止使用任何机器翻译模型或API
    - 句子为完整句子，至少包含4个以上汉字
  - 评价
    - 句对 BLEU 值

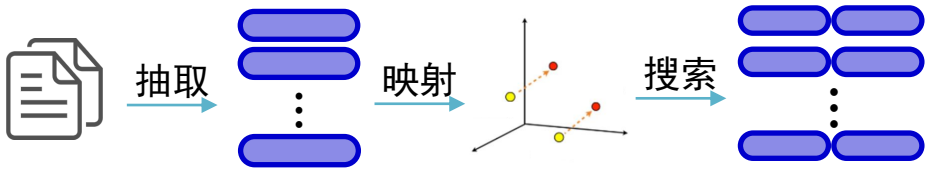
整体思路

- 无监督跨语言句对抽取任务 $D \rightarrow (S_{source}, S_{target})$



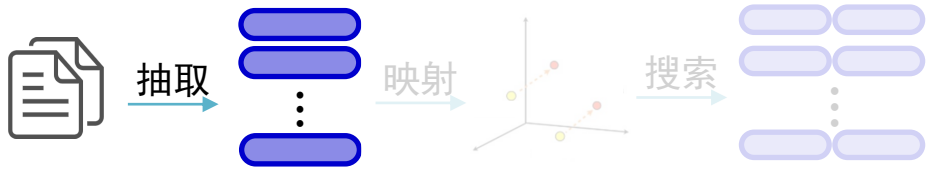
整体思路

- 从文档中抽取候选句子
  - $D \rightarrow S$
- 句子映射到语义空间
  - $S \rightarrow V$
- 在句子语义空间搜索最近邻
  - $(V_{query}, V_{dataset}) \rightarrow (S_{source}, S_{target})$



抽取候选句子

- 观察数据
  - 原始网页噪声多
  - 数据量大
- 方法
  - 数据清洗
  - 手工规则
- 候选句
  - 中文：1241279句
  - 意大利语：2922320句

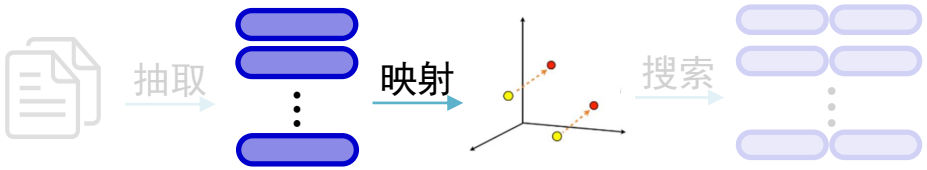



```
<!doctype html>
<html>
<head>
<script src="/js/m.js"></script>
<script type="text/javascript">var mi='/wap/yidalizhongwenxinwen/yihui/2019/1105/148325.html';m_qingtiancms_com(mi)</script>
<meta charset="utf-8" />
<title>时隔11年，罗马斗兽场门票涨价了，门票有效期也缩短了!_意大利新闻网</title>
<meta name="keywords" content="时隔,11年,罗马,斗兽场,门票,涨价,了,有效期," />
<meta name="description" content="（内容来自：意烩 oushitalia）11月1日起，罗马斗兽场的基本门票从之前的12欧元上涨至16欧元，并且有效期从之前的连续2日内有效，缩短为1日内有效。 罗马斗兽场的基本门票为3个景点联" />
<link rel="stylesheet" href="/templets/default/css_mubanzhijia_com/layout.css" type="text/css" />
```

Fig. 原始网页样例

映射句子到语义空间

- 难点
  - 候选句子数量大
  - 多语种映射到相同的语义空间



- 方法
  - 多语种语言模型
    - 基于 BERT 的语言模型
- 句子的特征向量
  - 512/768 维

Fig. Bert [2]

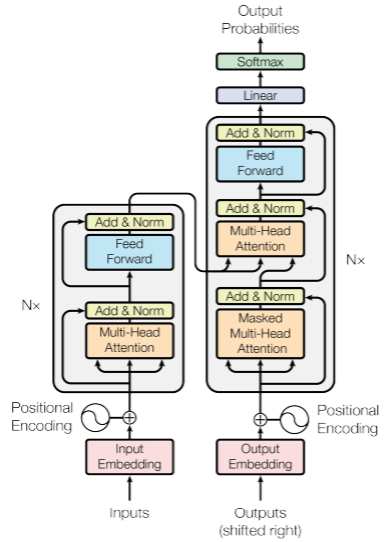
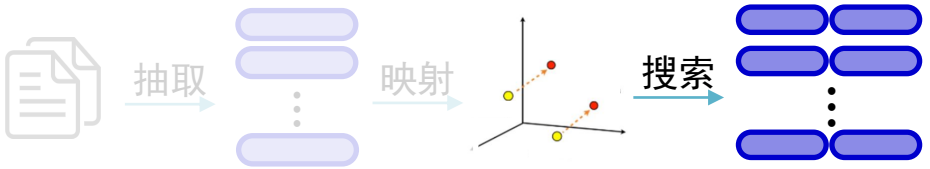


Fig. Transformer [1]

[1] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.
[2] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv*. 2018.

搜索语义空间最近邻

- 难点
  - 搜索复杂度高（暴力搜索M×N）



- 方法
  - Scann [1]
    - 量化
    - GPU 并行
  - 搜索用时：25分钟

$(V_{query}, V_{dataset}) \rightarrow (S_{source}, S_{target})$

V_{query}：中文候选句， 1241279句
V_{dataset}：意大利语候选句， 2922320句

[1] Guo Ruiqi, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. 2020. Accelerating Large-Scale Inference with Anisotropic Vector Quantization. In ICML.

大纲

- 团队
- 整体方案
- 后续优化
- 总结

优化

- 句子抽取
  - 抽取规则：句长，语种，正则，overlap，n-gram
- 特征映射
  - 尝试不同预训练模型（结构，预训练任务，预训练语料）
  - xlm-roberta[1]，m-USE[2]，LaBSE[3]
- 最近邻搜索
  - Partitioning(tree_leaves)
  - Scoring(brute-force/hashing, l2/cosine)
  - Rescoring: top-1, top-3, top-5

[1] Conneau, Alexis, et al. "Unsupervised cross-lingual representation learning at scale." arXiv preprint arXiv:1911.02116 (2019).
[2] Yang, Yinfei, et al. "Multilingual universal sentence encoder for semantic retrieval." arXiv preprint arXiv:1907.04307 (2019).
[3] Feng, Fangxiaoyu, et al. "Language-agnostic BERT Sentence Embedding." arXiv preprint arXiv:2007.01852 (2020).

结果

- 初赛

排名	参赛团队	分数	提交次数
1	肉蛋葱鸡	5034.2432	22
2	====baseline====	4294.36807	39
3	HNwaz8j8x	3613.77159	7

- 复赛

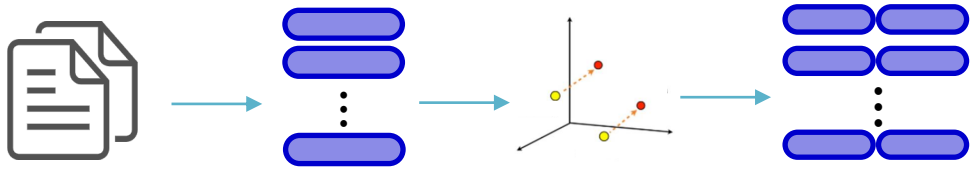
排名	参赛团队	分数	提交次数
1	肉蛋葱鸡	7562.11126	57
2	====baseline====	6694.19942	42
3	HNwaz8j8x	2394.87172	15

大纲

- 团队
- 整体方案
- 后续优化
- 总结

总结

- 理解赛题+观察数据
- 数据预处理
- 数据量大 → 并行、缓存
- 大规模预训练模型的潜力
- 阅读论文、开源项目（我们的代码将在1024节后开源）
- 比赛实践
- 快速迭代
  - 勤写测试
  - 解耦



Thank you.
Q&A