那些击溃了所有NLP系统的样本

原创 夕小瑶 夕小瑶的卖萌屋 2019-09-17

来自专辑
卖萌屋@自然语言处理

> 本文写作素材取自知乎文章
> https://zhuanlan.zhihu.com/p/55571643，作者mountain blue，知乎专栏《机器学习小知识》，快去关注一波~

非常重要的前言

无论你是PM还是QA还是java开发，请不要拿本文刺激你身边的NLP工程师，人生已经如此的艰难，有些事情就

击溃拼音标注系统篇

写给卖豆芽的对联，我想打印出拼音

长长长长长长长，长长长长长长长。

(solution: changzhangchangzhangchangchangzhangzhangchangzhangchangzhangzhangchang, zhangchangchangzhangchangzhangchangchangzhangchangchangzhangchangchangchang)

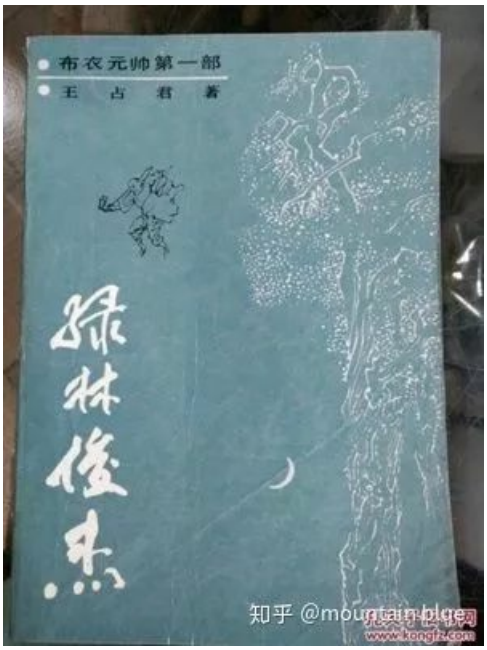
击溃词法分析系统（分词/POS/NER）篇

1. 来到杨过曾经生活过的地方，小龙女动情地说：“我也想过过过儿过过的生活。”
2. 看见西门吹雪点上了灯，叶孤城冷笑着说：“我也想吹吹吹雪吹过的灯”，然后就吹灭了灯。
3. 灭霸把美队按在地上一边摩擦一边给他洗脑，被打残的钢铁侠说：灭霸爸爸叭叭叭叭儿的在那叭叭啥呢
4. 来到儿子等校车的地方，邓超对孙俪说：“我也想等等等等等过的那辆车。”
5. 你也想犯范范范玮琪犯过的错吗
6. 碳碳键键能能否否定定律一
   （solution: 碳碳键/键能/能否/否定/定律一）

7. 对叙打击是一次性行为?

(solution: 对叙打击是一次性/行为)

8. 《绿林俊杰》



(ps：我家林俊杰做错了什么？为什么要绿他)

9. 书《无线电法国别研究》

(solution: 无线电法/国别/研究)

击溃句法分析系统篇

“宝鸡有一群怀揣着梦想的少年相信在牛大叔的带领下会创造生命的奇迹网络科技有限公司”

（ps：该公司全名长达39个字，还是一句主谓宾齐全的句子。宝鸡工商部门表示，该公司属合法注册，但名字太长不利于刻公章开发票）

击溃用户画像系统篇

我的微博用户名是“**一位友好的哥谭市民**”，你们给我生成的画像是啥？

(ps：对话是自下而上进行的)

我的画像不是*city: 哥谭*，而是

而是，的哥+姓谭！

击溃词义消歧系统篇

要去见投资人，出门时，发现车钥匙下面压了一张员工的小字条，写着“**老板，加油！**”，瞬间感觉好有温度，当时心里就泪奔了。心里默默发誓：我一定会努力的！车开了15分钟后，没油了。。。

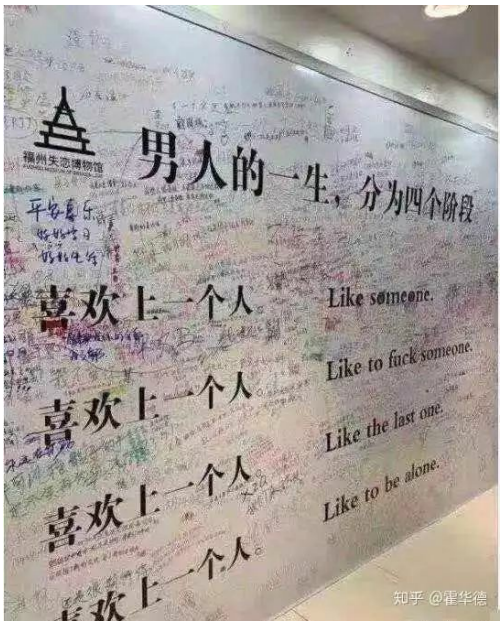
击溃指代消解系统篇

1. 他快抱不起儿子了，因为他太胖了

2. 宝宝的经纪人睡了宝宝的宝宝，宝宝不知道宝宝的宝宝是不是宝宝的亲生的宝宝，宝宝的宝宝为什么要这样对待宝宝！宝宝真的很难过！宝宝现在最担心的是宝宝的宝宝是不是宝宝的宝宝，如果宝宝的宝宝不是宝宝的宝宝那真是吓死宝宝了。

击溃机器翻译系统篇

1. 中->英 *（ps：这条素材取自知乎劝退男神“霍华德”，快去撩他！）*
“男人的一生，分为四个阶段：喜欢上一个人；喜欢上一个人；喜欢上一个人；喜欢上一个人。”



（ps：ELMo，BERT，XLNet，ERNIE也表示语义无法表示）

2. 英->中
"How can I help you?"

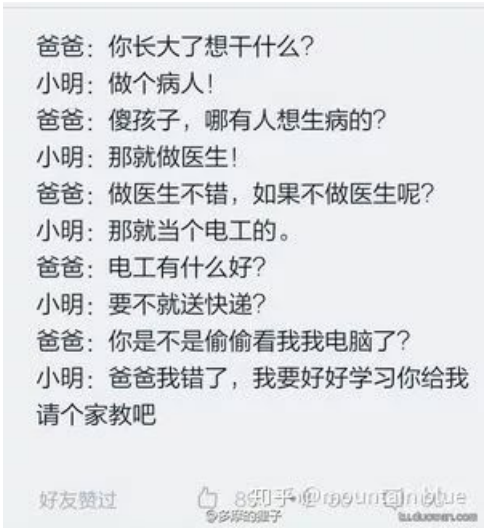
（ps：来自米国某酒店前台翻译机）

3. 文言->白话 *（拼音feature完全崩溃*
季姬寂，集鸡，鸡即棘鸡。棘鸡饥叽，季姬及箕稷济鸡。鸡既济，跻姬笈，季姬忌，急咭鸡，鸡急，继圾几，季姬急，即籍箕击鸡，箕疾击几伎，伎即畜，鸡叽集几基，季姬急极屐击鸡，鸡既殛，季姬激，即记《季姬击鸡记》。

击溃了fasttext词向量，无论中英

中文里面“大胜”和“大败”意思相同，刚发现英文里面也有类似的现象：valuable和invaluable都是表示非常有价值的意思

击溃了2050年的对话系统



(tip：对话领域分类是成熟对话系统的第一步，本题候选领域包括
A. 人生梦想类
B. 医学类
C. 教育类
D. 理工类
E. 爱情动作类 ）

本文后续

哼，哪一天小夕当了PM，就拿这篇帖子教训那些不听话的RD！尤其是那些做NLP的(^ -^)ノ