A NEW EVALUATION SYSTEM ON POS

面向词性标注评测的汉语易混淆词类研究

——从趋向动词-方位词的混淆展开

选题背景

Introduction

- 机器词性标注是自然语言处理诸多任务的基础，词性标注的正确与否直接关系到后续的句法分析、语义分析等下游任务。
- 机器词性标注的难点，是要解决易混淆词类的问题，尤其是兼类词在**易混淆结构**中的词性标注问题。

我看到鸟兽在树上山口那边乱作一团，接着便听到践踏芦苇的声音，不时传来枝条折断的嘎嘎声。

要千方百计保护好水稻田，提倡果树上山，鱼塘下滩。

选题背景

Introduction

树上山

选题背景

Introduction

我看到鸟兽在 **树上** 山口那边乱作一团，接着便听到践踏芦苇的声

选题背景

Introduction

要千方百计保护好水稻田，提倡果树 上山 ，鱼塘下滩。

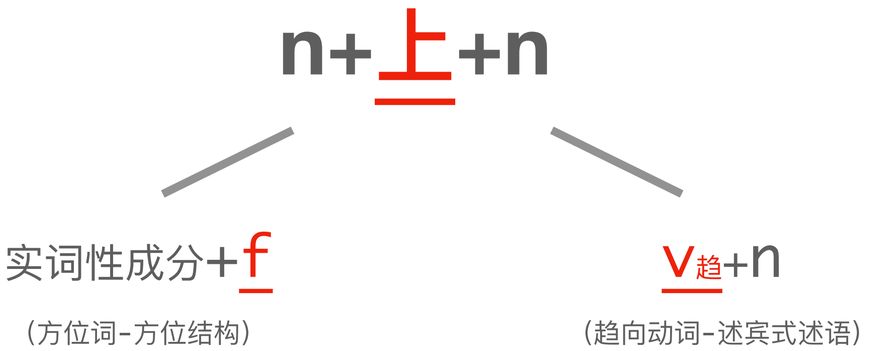
选题背景

Introduction

树上山

选题背景

Introduction



选题背景

Introduction

n+上+n

f-v的词性易混淆结构

选题背景

Introduction

f-v的词性易混淆结构

- 相同：两者都在空间关系及空间运动描述中起重要作用。鉴于空间结构同概念结构之间的映射关系、空间结构在意向图式中的重要地位，对f-v两者的研究一定程度上有助于研究人类的认知模型，推动机器的多模态认知处理。
- 不同：两者激活的意象图式有差异，扫描方式[①]也不同，这造成了两者在认知上的巨大差异。因此，区分他们是很有价值的。
  - 趋向动词主要激活路径模式（path）；方位词可以激活路径模式，同时也可以激活容器-内容图式（inside-outside\container-contained等）、方向图示（in-out\inside-outside\ front-back等）

① 扫描：认知科学上指在建构复杂场景时所做的认知处理。（Langacker, 1987；转引自张敏, 1998）

选题背景

Introduction

f-v的词性易混淆结构

- 不同：两者激活的意象图式有差异，扫描方式[①]也不同，这造成了两者在认知上的巨大差异。因此，区分他们是很有价值的。
  - 趋向动词主要激活路径模式（path）；方位词可以激活路径模式，同时也可以激活容器-内容图式（inside-outside\container-contained等）、方向图示（in-out\inside-outside\ front-back等）
  - 趋向动词是次第扫描(equential scanning)，方位词则是总括扫描(summary scanning)。对于同一个意象图式，次第扫描是依次处理不同时刻的每一个状态，得到的结果随着时间变化，如同一段电影；总括扫描则是将每个时刻都投射到更低维的空间，如同一张叠印照片。

王琦,郭锐.汉语趋向动词用作方向词现象初探[J].语言学论丛,2013(01):70-102+379-380.

选题背景

Introduction

f-v的词性易混淆结构

- 趋向动词是次第扫描(equential scanning)，方位词则是总括扫描(summary scanning)。对于同一个意象图式，次第扫描是依次处理不同时刻的每一个状态，得到的结果随着时间变化，如同一段电影；总括扫描则是将每个时刻都投射到更低维的空间，如同一张叠印照片。
- 具体到“上”这一个f-v兼类词：
  - 树/n 上/f：是总括扫描，识别的是空间要素(figure & background)之间的静态空间位置关系。
  - 往/p 上/f 爬/v：是总括扫描，识别的是figure相对于background运动时的整体轨迹。
  - 上/v 树/n：是次第扫描，识别的是figure相对于background的整个运动过程

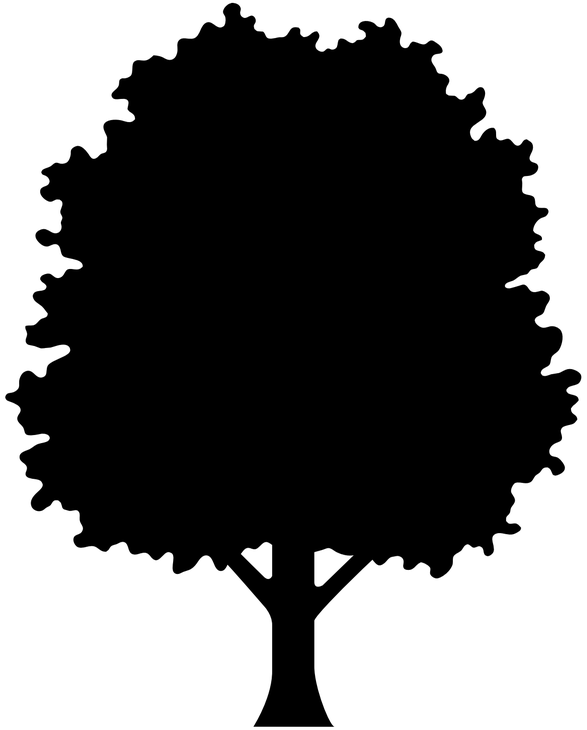
① 扫描：认知科学上指在建构复杂场景时所做的认知处理。（Langacker, 1987；转引自张敏, 1998）

选题背景

Introduction

f-v的词性易混淆结构

- 具体到“上”这一个f-v兼类词：
  - 树/n 上/f：是总括扫描，识别的是空间要素(figure & background)之间的静态空间位置关系。
  - 往/p 上/f 爬/v：是总括扫描，识别的是figure相对于background运动时的整体轨迹。
  - 上/v 树/n：是次第扫描，识别的是figure相对于background的整个运动过程



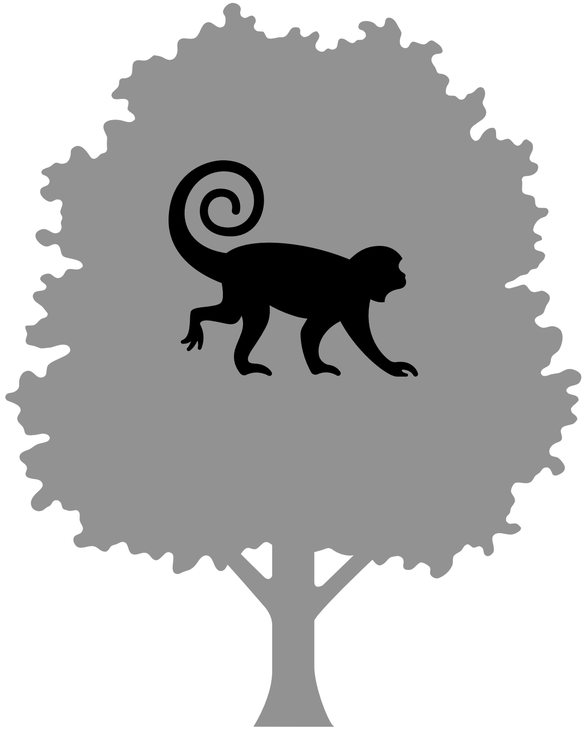
① 扫描：认知科学上指在建构复杂场景时所做的认知处理。（Langacker, 1987；转引自张敏, 1998）

选题背景

Introduction

f-v的词性易混淆结构

- 具体到“上”这一个f-v兼类词：
  - 树/n 上/f：是总括扫描，识别的是空间要素(figure & background)之间的静态空间位置关系。
  - 往/p 上/f 爬/v：是总括扫描，识别的是figure相对于background运动时的整体轨迹。
  - 上/v 树/n：是次第扫描，识别的是figure相对于background的整个运动过程



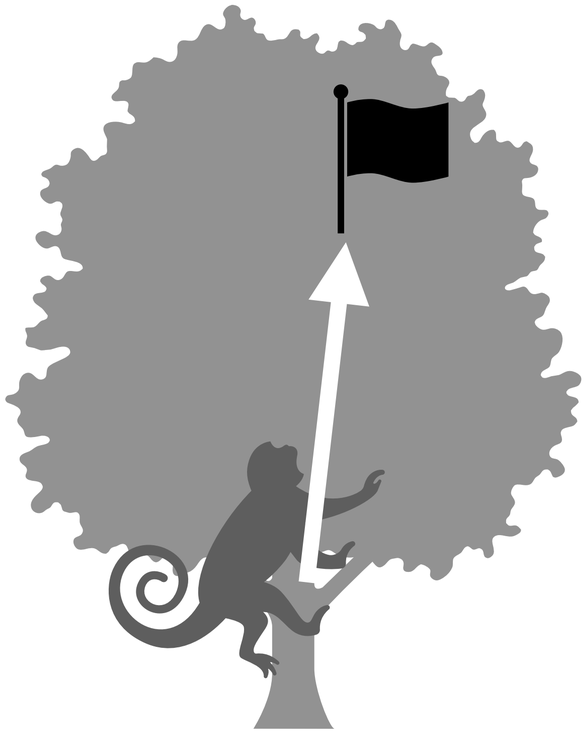
① 扫描：认知科学上指在建构复杂场景时所做的认知处理。（Langacker, 1987；转引自张敏, 1998）

选题背景

Introduction

f-v的词性易混淆结构

- 具体到“上”这一个f-v兼类词：
  - 树/n 上/f：是总括扫描，识别的是空间要素(figure & background)之间的静态空间位置关系。
  - 往/p 上/f 爬/v：是总括扫描，识别的是figure相对于background运动时的整体轨迹。
  - 上/v 树/n：是次第扫描，识别的是figure相对于background的整个运动过程



① 扫描：认知科学上指在建构复杂场景时所做的认知处理。（Langacker, 1987；转引自张敏, 1998）

选题背景

Introduction

f-v的词性易混淆结构

- 具体到“上”这一个f-v兼类词：
  - 树/n 上/f：是总括扫描，识别的是空间要素(figure & background)之间的静态空间位置关系。
  - 往/p 上/f 爬/v：是总括扫描，识别的是figure相对于background运动时的整体轨迹。
  - 上/v 树/n：是次第扫描，识别的是figure相对于background的整个运动过程

① 扫描：认知科学上指在建构复杂场景时所做的认知处理。（Langacker, 1987；转引自张敏, 1998）

选题背景

Introduction

n+上+n

f-v的词性易混淆结构

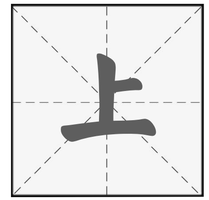
选题背景

Introduction

n+上+n

从f-v的词性易混淆结构展开

选题背景



Introduction

介詞

v趋+n
（趋向动词-述宾式述语）
上山 上车
上街
上水 上菜
上刺刀 上弦
上账
上年纪
上课

v+v趋+n+去/来
（趋向动词-可与宾语组合）
求上门
跑上山
找上门
冲上台
*走上前

v+v趋
（趋向动词-述补式补语）
锁上
选上
镀上

方位詞

p+f
（方位词-介宾式宾语）
在上
自上而下
从下到上
往上走
向上看

实词性成分+f
（方位词-方位结构中）
树上
桌上
在树上
在花钱上
到天上去

f
（方位词-对举主语/宾语）
上有老，下有小
欺上瞒下

指示詞

*zs+(m+)q
（指示词）
*上半年
*上半场
*上半城
*上册
*上集
*上一册
*上一集
*上一个

选题背景

Introduction

- 机器词性标注是自然语言处理诸多任务的基础，词性标注的正确与否直接关系到后续的句法分析、语义分析等下游任务。
- 机器词性标注的难点，是要解决易混淆词类的问题，尤其是兼类词在易混淆结构中的词性标注问题。从数据上看，汉语中不到6%的兼类词造成了机器词性标注46~64%的错误。
- 方位词和趋向动词的混淆，关系到汉语空间认知理解的问题，在机器的多模态认知处理、机器类人化研究等领域有先导性价值。

研究进度

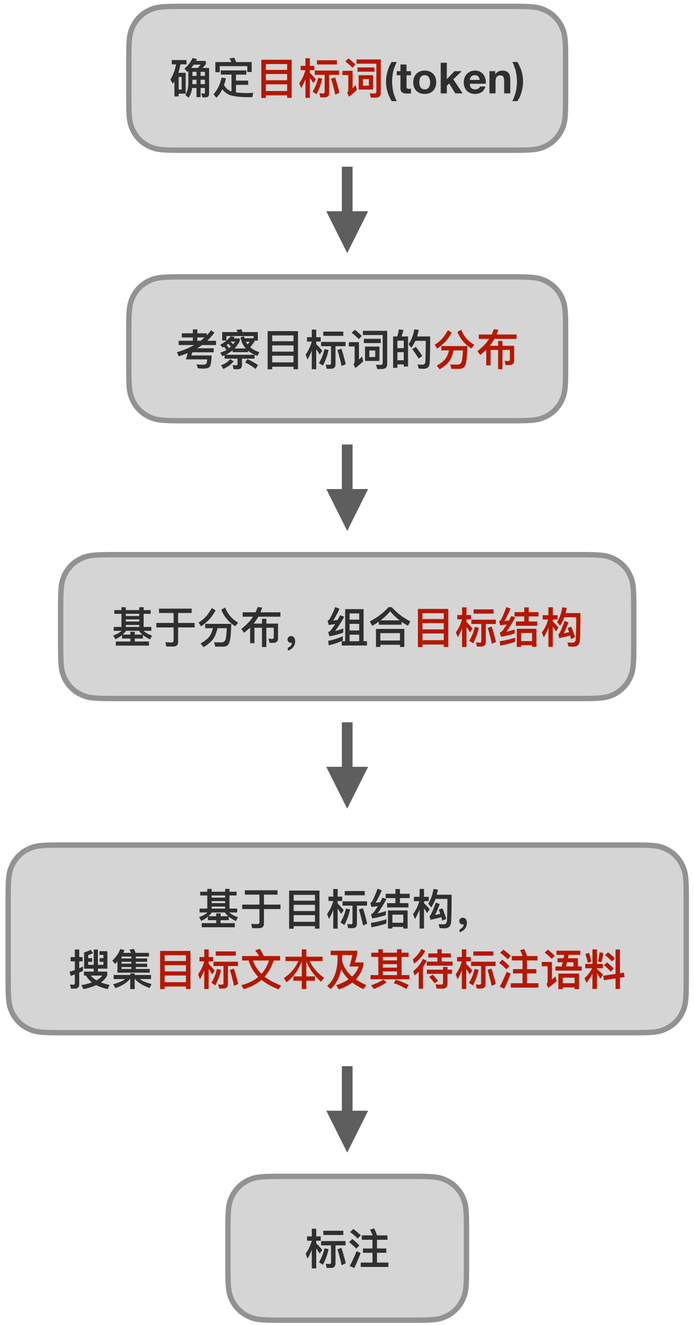
Completed

研究进度

Completed

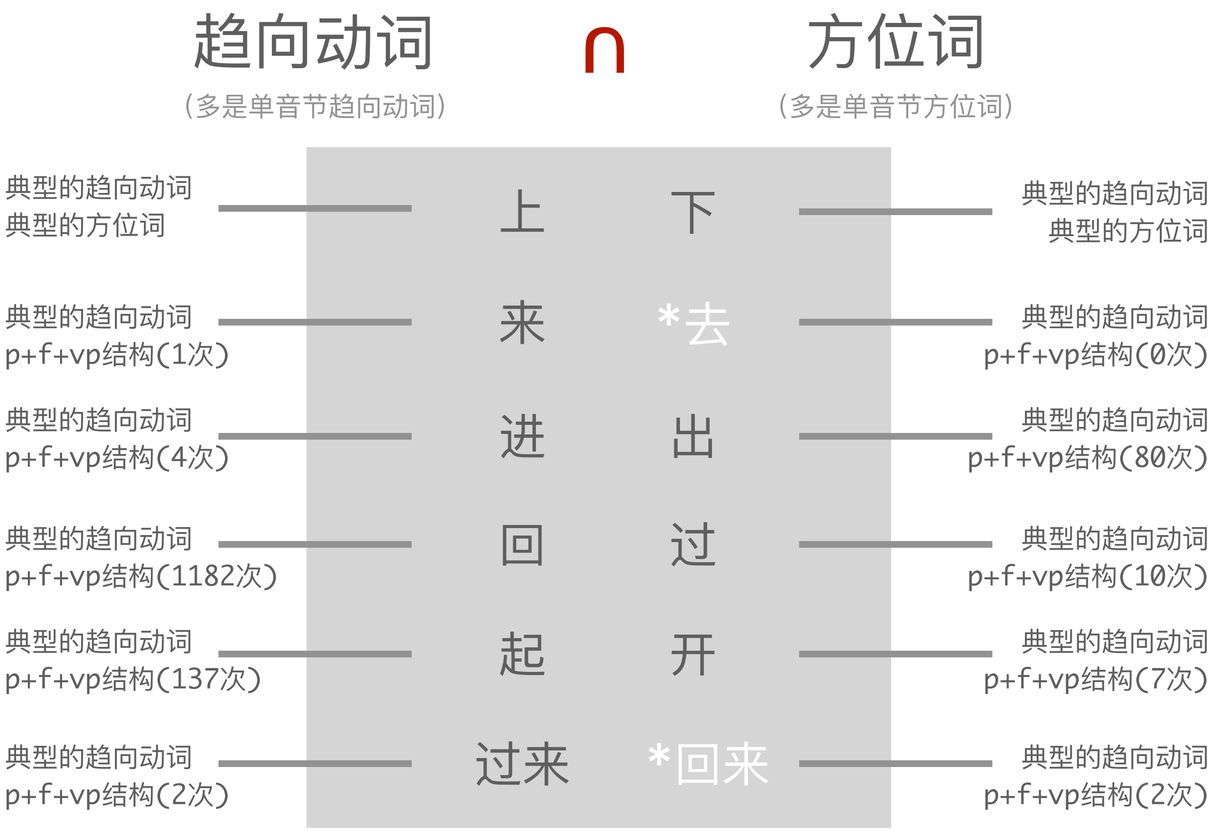
采用描写、搜集、标注的方法

产生试题集



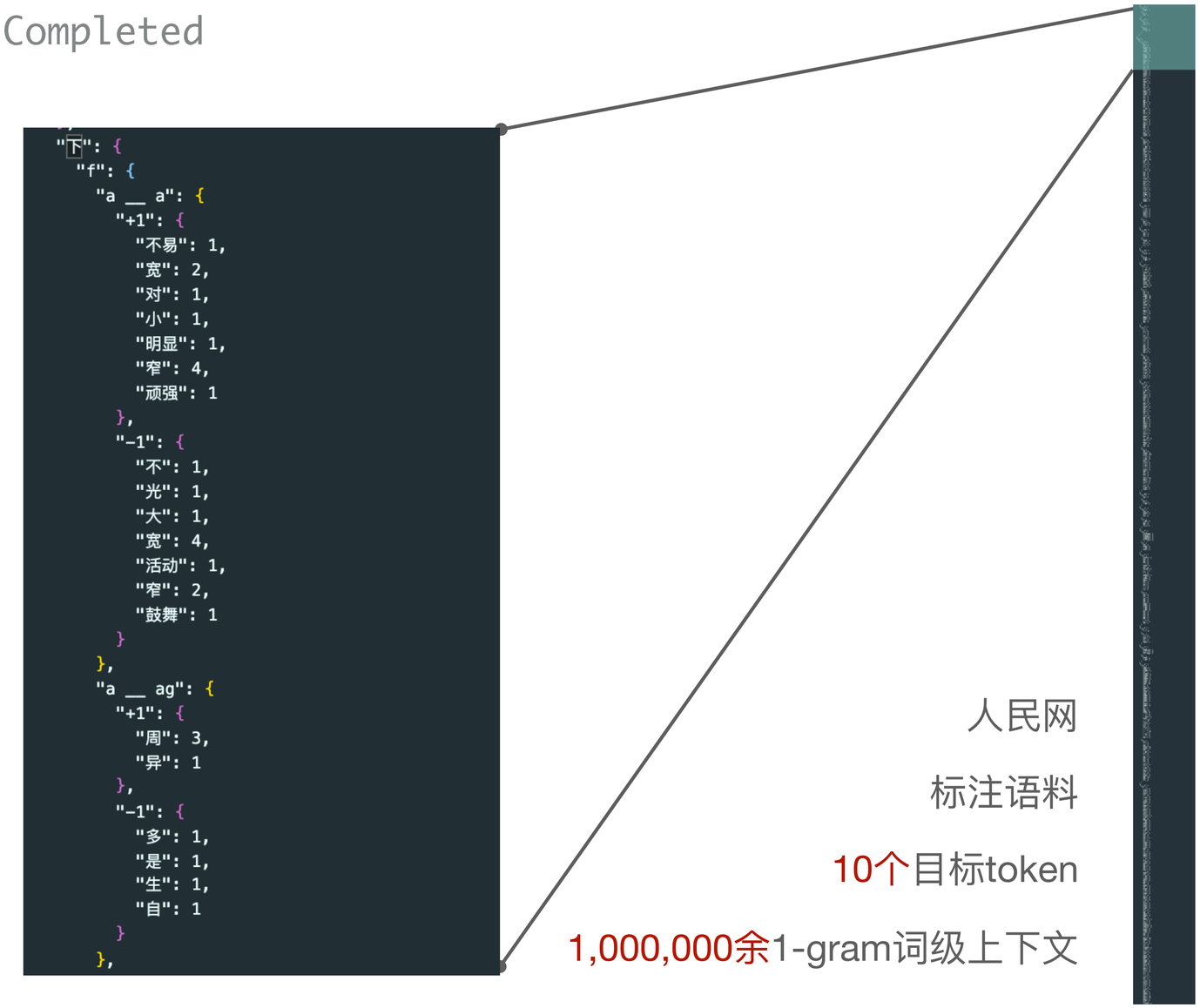
研究进度

Completed



研究进度

Completed



人民网

标注语料

10个目标token

1,000,000余1-gram词级上下文

研究进度

Completed

1,000,000余1-gram词级上下文

20,000余条具体分布

1,000余个有混淆可能性的结构

190余个目标结构

23行正则表达式

研究进度

Completed

1,000,000余1-gram词级上下文

20,000余条具体分布

1,000余个有混淆可能性的结构

190余个目标结构

23行正则表达式

研究进度

Completed

```
"下": [
  "entity 下 a entity",
  "entity 下 entity",
  "entity 下 m a entity",
  "entity 下 m entity",
  "entity 下 m q a entity",
  "entity 下 m q entity",
  "entity 下 q a entity",
  "entity 下 q entity",
  "entity 下 下 entity",
  "p entity 下 a entity",
  "p entity 下 entity",
  "p entity 下 m entity",
  "p entity 下 m q a entity",
  "p entity 下 m q entity",
  "p entity 下 q entity",
  "p entity 下 下 entity",
  "p 下 a entity",
  "p 下 entity",
  "p 下 m entity",
  "p 下 m q entity",
  "p 下 q entity",
  "一 下 a entity",
  "一 下 entity",
  "一 下 下 a 的 entity",
  "一 下 下 entity",
  "下(reduplication)"
],
"出": [
  "一 出 a entity",
  "一 出 entity",
```

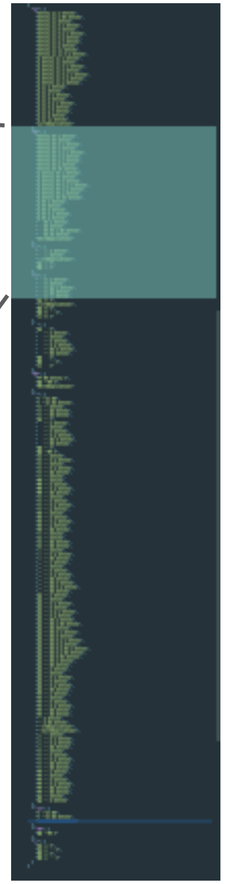
1,000,000余1-gram词级上下文

20,000余条具体分布

1,000余个有混淆可能性的结构

190余个目标结构

23行正则表达式



确定目标词(token)

↓

考察目标词的分布

↓

基于分布，组合目标结构

↓

基于目标结构，
搜集目标文本及其待标注语料

↓

标注

研究进度

Completed

1,000,000余1-gram词级上下文

20,000余条具体分布

1,000余个有混淆可能性的结构

190余个目标结构

23行正则表达式

```
    # r'entity 上 v',
],
'下': [
    r'((p entity)|p|entity)( 下)+( m)?( q( 的)?)?( a( 的)?)? entity',
    r'一( 下)+( a( 的)?)? entity',
],
'进': [
    r'往( 进)+( 一)? v',
```

确定目标词(token)

考察目标词的分布

基于分布，组合目标结构

基于目标结构，
搜集目标文本及其待标注语料

标注

研究进度

Completed

```
'上': [
    r'((p entity)|p|entity)( 上)+( m)?( q( 的)?)?( a( 的)?)? entity',
    # r'entity 上 v',
],
```


```
"entity 上 a entity": [
  200
],
"entity 上 a 的 entity": [
  3
],
"entity 上 entity": [
  11767
],
"entity 上 m a entity": [
  59
],
"entity 上 m entity": [
  648
],
"entity 上 m q a entity": [
  43
],
"entity 上 m q entity": [
  641
],
"entity 上 q a entity": [
  1
],
"entity 上 q entity": [
  30
],
"entity 上 上 entity": [
  22
],
"entity 上 上 m q entity": [
  2
],
"上(reduplication)": [
  1447
],
```

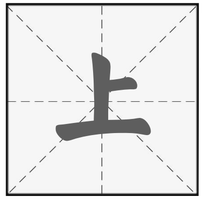

确定目标词(token)

考察目标词的分布

基于分布，组合目标结构

基于目标结构，
搜集目标文本及其待标注语料

标注



```
'上': [
    r'((p entity)|p|entity)( 上)+( m)?( q( 的)?)?( a( 的)?)? entity',
    # r'entity 上 v',
],
```

目标结构

```
"entity 上 a entity": [
  200
],
"entity 上 a 的 entity": [
  3
],
"entity 上 entity": [
  11767
],
"entity 上 m a entity": [
  59
],
"entity 上 m entity": [
  648
],
"entity 上 m q a entity": [
  43
],
"entity 上 m q entity": [
  641
],
"entity 上 q a entity": [
  1
],
"entity 上 q entity": [
  30
],
"entity 上 上 entity": [
  22
],
"entity 上 上 m q entity": [
  2
],
"上(reduplication)": [
  1447
```

context实例（简化版）

从文学界辐射到社会上一般阅读者
该类板块股票将带领大市上新台阶
朗平的音义是海面上平静的浪
这时巷口上灯光一亮
阳城乡一位负责人上县开会时
是当代文学史上许多重要事件的见证人
促成中国场地赛车运动上一个新台阶
大街上不少女性都穿起了短衫和短裙
对北京武术运动上一个台阶充满信心
旗杆上七丛白毛迎风飘扬
叫锯碗的钉上几个小铜钉
头上一顶青缎小帽
要在你们北影上一部电影
但看到壁上种种奇妙招数
鼻中闻到她身上阵阵幽香
母亲上个月十五做了一梦
我就在这条街上上班
争取把上届所得的铜牌上上金色
在乐器和谱架上插上中国国旗

目标文本及其自动词性标注

社会/n 上/v\f 一般/a 阅读者/n
大市/n 上/v 新/a 台阶/n
海面/n 上/v\f 平静/a 的/u 浪/n
巷口/n 上/v\f 灯光/n
负责人/n 上/v 县/n
文学史/n 上/v\f 许多/m 重要/a 事件/n
运动/v 上/v 一个/m 新/a 台阶/n
大街/n 上/v\f 不少/m 女性/n
运动/v 上/v 一个/m 台阶/n
旗杆/n 上/v\f 七/m 丛/q 白/a 毛/n
钉/v 上/v 几/m 个/q 小/a 铜钉/n
头/n 上/v\f 一/m 顶/q 青缎/n
北影/j 上/v 一/m 部/q 电影/n
壁/n 上/v\f 种种/q 奇妙/a 招数/n
身上/s(f) 阵阵/q 幽香/n
母亲/n 上个月/t(r)
街上/s(f) 上班/v(v)
铜牌/n 上上/v 金色/n
上/v\f 插/v 上/v

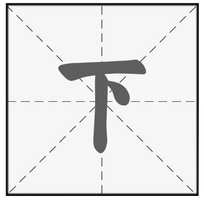
确定目标词(token)

考察目标词的分布

基于分布，组合目标结构

基于目标结构，
搜集目标文本及其待标注语料

标注



```
'下': [
    r'((p entity)|p|entity)( 下)+( m)?( q( 的)?)?( a( 的)?)? entity',
    r'一( 下)+( a( 的)?)? entity',
],
```


```
"entity 下 a entity": [
  131,
],
"entity 下 entity": [
  6607,
],
"entity 下 m a entity": [
  49,
],
"entity 下 m entity": [
  333,
],
"entity 下 m q a entity": [
  11,
],
"entity 下 m q entity": [
  196,
],
"entity 下 q a entity": [
  2,
],
"entity 下 q entity": [
  19,
],
"entity 下 下 entity": [
  3,
],
"p entity 下 a entity": [
  16,
],
"p entity 下 entity": [
  566,
],
"p entity 下 m entity": [
  22,
],
"p entity 下 m q a entity": [
  2,
],
"p entity 下 m q entity": [
  23,
```


考察标的　　参考示例

下

【方位词f】

他抬起一个脚丫踩在我脸上，用力 往/p 下/f 一/m 踹/v ，我便摔回池中。

【指示词(暂计入代词)r】

行乞完，他们又必然会到官府赖求，再盖一个官印，成为 向/p 下/r 一/m 站/n 行乞的“签证”。

【动词v】

这两人 一/d 下/v 山/n ，群雄问起陆菲青别来情形。

【量词q】

主意一定，郑涨钱托人照看 一/m 下/q 山林/n ，带着五岁的平波去老方桥刘桂英家。

确定目标词(token)

考察目标词的分布

基于分布，组合目标结构

基于目标结构，
搜集目标文本及其待标注语料

标注

研究进度

Completed

考察标的　参考示例

上

【方位词f】

天色暗下，他叫堂房去街上买回一碗肉丝面，然后和堂房闲聊几句，假似无意地问：楼/n 上/f 客人/n | 楼上/s 客人/n 今晚是否都在?

【动词v】

堂房轻轻叩一下他的房门："杭州 客人/n 上/v 楼/n | 客人/n 上楼/v 了。"

今后一年内，上海将以平均每月 上/v 一/m 个/q 项目/n 的速度在这一地区开办一系列独资生产企业。

【指示词(暂计入代词)r】

全队精神振奋，士气高昂，互相鼓励，做完了 上/r 一/m 个/q 项目/n，就积极准备下一个项目的动作。

考察标的　参考示例

下

【方位词f】

他抬起一个脚丫踩在我脸上，用力 往/p 下/f 一/m 踹/v，我便摔回池中。

【指示词(暂计入代词)r】

行乞完，他们又必然会到官府赖求，再盖一个官印，成为 向/p 下/r 一/m 站/n 行乞的"签证"。

【动词v】

这两人 一/d 下/v 山/n，群雄问起陆菲青别来情形。

【量词q】

主意一定，郑涨钱托人照看 一/m 下/q 山林/n，带着五岁的平波去老方桥刘桂英家。

考察标的　参考示例

来

【动词v】

今年/t 来/v 北京/ns 的/u 外国留学生计划将增加1万人,增加人数将是目前在京留学生人数的三分之一。

【方位词f】

的确，近年/t 来/f 北京/ns 的/u 城市面貌用"日新月异"来形容毫不夸张。

【助词u】

孩子上学，也就那 十/m 来/u 年/q 的时光。

标出这个片段的正确切词和词性 注意：
1．需要标注者考虑目标token不同的切分颗粒度，给出尽可能多的正确答案。不同答案之间用"|"隔开
2．仅用考虑目标token不同的切分颗粒度即可，不用考虑其他token的切分颗粒度。
3．如果不宜将目标token独立切分为一个清晰的语言单位，则

研究进度

Completed

考察标的　参考示例

回

【动词v】

星子 一/d 回/v 家/n ，次日就去找粞。

【量词q】

北伐以前，孙中山先生曾特派代表送了个第一师长的委任状来，请了 一/m 回/q 客/n ，送了两千元路费。

【方位词(仅表方向)f】

沿着原来的北上长征路线 往/p 回/f 走/v 。

标出这个片段的正确切词和词性 注意：
1．需要标注者考虑目标token不同的切分颗粒度，给出尽可能多的正确答案。不同答案之间用“|”隔开
2．仅用考虑目标token不同的切分颗粒度即可，不用考虑其他token的切分颗粒度。
3．如果不宜将目标token独立切分为一个清晰的语言单位，则应该弃用这条数据。例如：“往来”、“部下”等词不宜分成“往/v 来/v”、“部/n 下/f”，应该弃用；“墙上”可以标为“墙/n 上/f”，可以保留。

考察标的　参考示例

出

【动词v】

我们看的时候涕泪交加，可 一/d 出/v 影院/n 就恍若隔世。

【量词q】

我像亲自串演了 一/m 出/q 人间/n 的悲剧，心头浸蚀了无名的怅惘。

【方位词(仅表方向)f】

血从后背“汩汩” 往/p 出/f 冒/v 。

标出这个片段的正确切词和词性 注意：
1．需要标注者考虑目标token不同的切分颗粒度，给出尽可能多的正确答案。不同答案之间用“|”隔开
2．仅用考虑目标token不同的切分颗粒度即可，不用考虑其他token的切分颗粒度。
3．如果不宜将目标token独立切分为一个清晰的语言单位，则应该弃用这条数据。例如：“往来”、“部下”等词不宜分成“往/v 来/v”、“部/n 下/f”，应该弃用；“墙上”可以标为“墙/n 上/f”，可以保留。

考察标的　参考示例

进

【动词v】

【方位词(仅表方向)f】

你放屁！宋大人堂堂四品提刑官传一个平民百姓还须亲自来请？姓刁的敢摆那么大的谱？来呀众弟兄随我一起 往/p 进/f 冲/v ！

标出这个片段的正确切词和词性 注意：
1．需要标注者考虑目标token不同的切分颗粒度，给出尽可能多的正确答案。不同答案之间用“|”隔开
2．仅用考虑目标token不同的切分颗粒度即可，不用考虑其他token的切分颗粒度。
3．如果不宜将目标token独立切分为一个清晰的语言单位，则应该弃用这条数据。例如：“往来”、“部下”等词不宜分成“往/v 来/v”、“部/n 下/f”，应该弃用；“墙上”可以标为“墙/n 上/f”，可以保留。

考察标的　参考示例

开

【动词v】

【方位词(仅表方向)f】

贺人杰赶紧 往/p 开/f 一/m 让/v ，蔡天化回手一拳，出其不意，认定黄天霸肩背上一击。

标出这个片段的正确切词和词性 注意：
1．需要标注者考虑目标token不同的切分颗粒度，给出尽可能多的正确答案。不同答案之间用“|”隔开
2．仅用考虑目标token不同的切分颗粒度即可，不用考虑其他token的切分颗粒度。
3．如果不宜将目标token独立切分为一个清晰的语言单位，则应该弃用这条数据。例如：“往来”、“部下”等词不宜分成“往/v 来/v”、“部/n 下/f”，应该弃用；“墙上”可以标为“墙/n 上/f”，可以保留。

研究进度

Completed

考察标的　参考示例

过去

【时间词t】

老年人一旦遇到不如意的事或挫折，便总喜欢 拿/v 过去/t 的/u 经历/n 来对比，往往沉湎于往事而不能自拔，这种心理现象在心理学上称为回归心理。

【动词v】

在各种各样成堆的商品前，中国人和越南人一边指手画脚，一边用计算器讨价还价，不管从中国 拿/v 过去/v 的/u 商品/n 是什么，肯定都会围上一大圈越南人来找你讲价。

标出这个片段的正确切词和词性 注意：
1．需要标注者考虑目标token不同的切分颗粒度，给出尽可能多的正确答案。不同答案之间用“|”隔开
2．仅用考虑目标token不同的切分颗粒度即可，不用考虑其他token的切分颗粒度。
3．如果不宜将目标token独立切分为一个清晰的语言单位，则应该弃用这条数据。例如：“往来”、“部下”等词不宜分成“往/v 来/v”、“部/n 下/f”，应该弃用；“墙上”可以标为“墙/n 上/f”，可以保留。

考察标的　参考示例

过来

【动词v】

【方位词(仅表方向)f】

那人站在门口，慢慢地朝四下张望。终于望见了站在路当中的我，便不再望了。慢慢 往/p 过来/f 走/v 。

标出这个片段的正确切词和词性 注意：
1．需要标注者考虑目标token不同的切分颗粒度，给出尽可能多的正确答案。不同答案之间用“|”隔开
2．仅用考虑目标token不同的切分颗粒度即可，不用考虑其他token的切分颗粒度。
3．如果不宜将目标token独立切分为一个清晰的语言单位，则应该弃用这条数据。例如：“往来”、“部下”等词不宜分成“往/v 来/v”、“部/n 下/f”，应该弃用；“墙上”可以标为“墙/n 上/f”，可以保留。

考察标的　参考示例

过

【助词u】

是的，去年年底从徐州到蚌埠，我 走/v 过/u 津浦路/n | 走过/v 津浦路/n 。记得那时为了避免敌机轰炸趁夜才能开车，多半是载运难民同军队的。

【动词v】

走/v 过/v 楼门/n | 走过/v 楼门/n 时，她又蜷成一团，把我的脑袋整个包住。

亚里士多德认为国家的产生不仅仅是为了生活，而是为了 过/v 好/a 的/u 生活/n ，优良的生活。

【副词d】

曹先生并不是吃“磁力学饭”的人，我们也就无须对他提出 过/d 高/a 的/u 要求/n 。

【方位词(仅表方向)f】

他一扭头，眼睛正好看见了那位还没找到座位的老太太，便抬起手杖，拨了拨坐在对面的一个和自己年龄相仿的人，并且把头 往/p 过/f 一/m 偏/v ，示意他站起来，把屁股下面的座位让出来。

【前接成分h】

标出这个片段的正确切词和词性 注意：
1．需要标注者考虑目标token不同的切分颗粒度，给出尽可能多的正确答案。不同答案之间用“|”隔开
2．仅用考虑目标token不同的切分颗粒度即可，不用考虑其他token的切分颗粒度。
3．如果不宜将目标token独立切分为一个清晰的语言单位，则应该弃用这条数据。例如：“往来”、“部下”等词不宜分成“往/v 来/v”、“部/n 下/f”，应该弃用；“墙上”可以标为“墙/n 上/f”，可以保留。

研究进度

Completed

研究进度

Completed

https://annotator-node-1300722527.cos.ap-beijing.myqcloud.com/index.html#/

研究进度

Completed

1. 首先，检查考察标的：
   1. 自动选定的这个片段，是否值得拎出来作为考察标的？如果：
      1. 自动分词与词性标注没有出错
      2. 考察标的与预期考察的评测项目不一致
      3. 标注者认为没有考察的必要

      必须至少满足上面三个条件中的两个，才能弃用这条数据。
   2. 如果标注者认为需要调整考察标的，可以进行调整
2. 然后，标出这个片段的正确切词和词性

   注意：

      1. 需要标注者考虑目标token不同的切分颗粒度，给出尽可能多的正确答案。不同答案之间用“|”隔开
      2. 仅用考虑目标token不同的切分颗粒度即可，不用考虑其他token的切分颗粒度。
      3. 如果不宜将目标token独立切分为一个清晰的语言单位，则应该弃用这条数据。例如：“往来”、“部下”等词不宜分成“往/v来/v”、“部/n 下/f”，应该弃用；“墙上”可以标为“墙/n 上/f”，可以保留。
3. 最后，检查目标结构是否正确，如果错误请修改
   1. 如果目标token在目标文本属于重复类型，则标reduplication
   2. 如果不是reduplication，则判断最小颗粒度的切分符合哪个目标结构。如果目标结构不在清单中，那么可以添加，添加须遵循以下要求：
      1. 添加的结构必须包含目标Token
      2. 添加的结构必须使用空格进行切分
      3. 上一步中最小颗粒度的切词，必须符合添加的结构；如果不符合的话，应该弃用这条数据。

研究进度

Completed

研究进度

Completed

研究进度

Completed

遇到的问题

Problems

遇到的问题

Problems

- P1：每个目标结构中，目标token的词性分布不均。并且很多易混淆结构更多是由词带来的，而不是由词性带来的。（例如：上一个项目/上一个台阶 vs　上 m q n）
  - S：1. 需要更细致地区分目标结构，尽可能具体到词。2. 标注后要做总结，尽可能总结出模板，便于生成。
- P2：一方面待标注数据规模较大（35775条数据待标注）；另一方面又有许多歧义结构没有搜索到
  - S：1. 需要更细致地区分目标结构。2. 需要改进搜索算法。3. 扩大语料规模。

下一步

Todo

- Todo：
  - 1. 组织标注。
  - 2. 在正式开始标注之前，细化目标结构（细化到上下文具体的词），尽可能均匀数据分布，提高数据质量，减少数据数量。
  - 3. 标注完成后要总结生成模板，或精进搜索模板。
  - 4. 跑基线。

A NEW EVALUATION SYSTEM ON POS

面向词性标注评测的汉语易混淆词类研究

——从趋向动词-方位词的混淆展开