

NLP – Natural language processing

自然语言处理

让 PM 全面理解NLP

自然语言处理 – NLP

核心任务

自然语言理解

自然语言生成

核心步骤

分词

词干提取
词形还原

词性标注

命名实体识别



easyAI – 产品经理的 AI 学习库

<https://easyai.tech>

下面所有内容均来自「[easyAI - 产品经理的 AI学习库](#)」。为了方便大家阅读和保存，整理成了 PDF 专题的形式。

如果觉得内容不错，可以访问我们的网站 [easyAI - 产品经理的 AI 学习库](#) 查看更多关于人工智能的科普文章。

也可以关注我们的公众号：产品经理的 AI 知识库（微信号：[easyai-tech](#)）



目录：

1. 自然语言处理
2. 自然语言理解
3. 自然语言生成
4. 分词
5. 词干提取、词形还原
6. 词性标注
7. 命名实体识别

NLP全景图

我们先通过一张长图来全面了解一下深度学习，如果想下载长图，可以访问这个链接-「[下载链接](https://easyai.tech)」

非技术一图看懂 NLP

产品经理的人工智能学习库

<https://easyai.tech>

详情访问: <https://easyai.tech>

1 什么是自然语言处理 – NLP



在人工智能出现之前，机器智能处理结构化的数据（例如 Excel 里的数据）。但是网络中大部分的数据都是非结构化的，例如：文章、图片、音频、视频..

为了能够分析和利用这些文本信息，我们就需要利用 NLP 技术，让机器理解这些文本信息，并加以利用。

NLP 就是人类和机器之间沟通的桥梁！

自然语言就是大家平时在生活中常用的表达方式，大家平时说的「讲人话」就是这个意思。

详情访问: <https://easyai.tech>

2 NLP 的 2 大核心任务

NLP的 2 个核心任务

NLU

NLG



自然语言理解（NLU）就是希望机器像人一样，具备正常人的语言理解能力，由于自然语言在理解上有很多难点(下面详细说明)，所以 NLU 是至今还远不如人类的表现。

NLG 是为了跨越人类和机器之间的沟通鸿沟，将非语言格式的数据转换成人类可以理解的语言格式，如文章、报告等。

详情访问：<https://easyai.tech>

3

NLP 的 5 个难点

- ① 语言是没有规律的，或者说规律是错综复杂的。
- ② 语言是可以自由组合的，可以组合复杂的语言表达。
- ③ 语言是一个开放集合，可以发明创造一些新的表达方式。
- ④ 语言需要联系到实践知识，有一定的知识依赖。
- ⑤ 语言的使用要基于环境和上下文。



详情访问：<https://easyai.tech>

4

NLP 的 4 个典型应用



情感分析



聊天机器人





语音识别



机器翻译

详情访问: <https://easyai.tech>

5 NLP 的 2 种实现方式

NLP 可以使用传统的机器学习方法来处理, 也可以使用深度学习的方法来处理。

2 种不同的途径也对应着不同的处理步骤。详情如下:



详情访问: <https://easyai.tech>

5 英文语料预处理的 6 个步骤

- 1 分词 – Tokenization
- 2 词干提取 – Stemming
- 3 词形还原 – Lemmatization
- 4 词性标注 – Parts of Speech tagging
- 5 命名实体识别 – NER
- 6 分块 – Chunking

详情访问: <https://easyai.tech>

6 中文语料预处理的 4 个步骤



- 1 中文分词 – Chinese Word Segmentation
- 2 词性标注 – Parts of Speech tagging
- 3 命名实体识别 – NER
- 4 去除停用词



easyAI – 产品经理的 AI 知识库

<https://easyai.tech>



公众号: easyai-tech

自然语言处理 - NATURAL LANGUAGE
PROCESSING | NLP



一文看懂自然语言处理 – NLP

easyai

网络上有海量的文本信息，想要处理这些非结构化的数据就需要利用 NLP 技术。

本文将介绍 NLP 的基本概念，2大任务，4个典型应用，5个难点和6个实践步骤。

NLP 为什么重要？

“语言理解是人工智能领域皇冠上的明珠”

比尔·盖茨

在人工智能出现之前，机器智能处理结构化的数据（例如 Excel 里的数据）。但是网络中大部分的数据都是非结构化的，例如：文章、图片、音频、视频...



结构化数据

Excel 数据库



非结构化数据

文本 图片 视频

easyai

在非结构数据中，文本的数量是最多的，他虽然没有图片和视频占用的空间大，但是他的信息量是最大的。

为了能够分析和利用这些文本信息，我们就需要利用 NLP 技术，让机器理解这些文本信息，并加以利用。

什么是自然语言处理 - NLP

每种动物都有自己的语言，机器也是！

自然语言处理（NLP）就是在机器语言和人类语言之间沟通的桥梁，以实现人机交流的目的。

人类通过语言来交流，狗通过汪汪叫来交流。机器也有自己的交流方式，那就是数字信息。



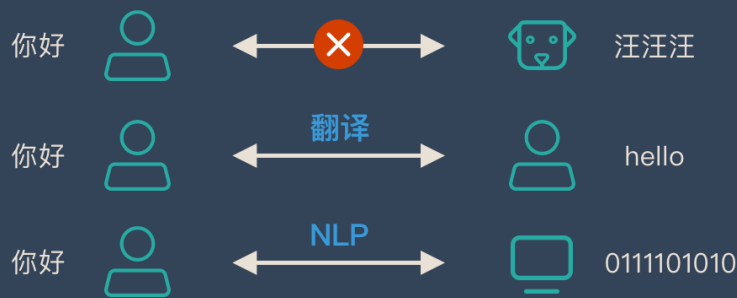
不同的语言之间是无法沟通的，比如说人类就无法听懂狗叫，甚至不同语言的人类之间都无法直接交流，需要翻译才能交流。

而计算机更是如此，为了让计算机之间互相交流，人们让所有计算机都遵守一些规则，计算机的这些规则就是计算机之间的语言。

既然不同人类语言之间可以有翻译，那么人类和机器之间是否可以通过“翻译”的方式来直接交流呢？

NLP 就是人类和机器之间沟通的桥梁！

NLP就是人类和机器之间沟通的桥梁



easyai

为什么是“自然语言”处理？

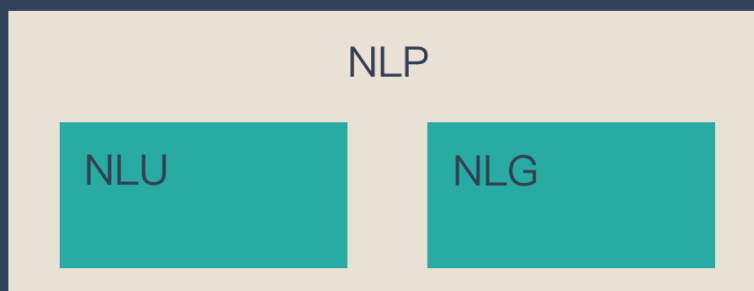
自然语言就是大家平时在生活中常用的表达方式，大家平时说的「讲人话」就是这个意思。

自然语言：我背有点驼(非自然语言：我的背部呈弯曲状)

自然语言：宝宝的经纪人睡了宝宝的宝宝（微博上这种段子一大把）

NLP 的2大核心任务

NLP 的2个核心任务



easyai

NLP 有2个核心的任务：

1. 自然语言理解 - NLU | NLI
2. 自然语言生成 - NLG

自然语言理解 - NLU | NLI

自然语言理解就是希望机器像人一样，具备正常人的语言理解能力，由于自然语言在理解上有很多难点(下面详细说明)，所以 NLU 是至今还远不如人类的表现。



自然语言理解就是希望机器像人一样，具备正常人的语言理解能力

easyai

自然语言理解的5个难点：

1. 语言的多样性
2. 语言的歧义性
3. 语言的鲁棒性
4. 语言的知识依赖
5. 语言的上下文

想要深入了解NLU，可以看看这篇文章《[一文看懂自然语言理解-NLU（基本概念+实际应用+3种实现方式）](#)》

自然语言生成 - NLG

NLG – 将非语言格式的数据转换成人类可



easyai

NLG 是为了跨越人类和机器之间的沟通鸿沟，将非语言格式的数据转换成人类可以理解的语言格式，如文章、报告等。

NLG 的6个步骤：

1. 内容确定 - Content Determination
2. 文本结构 - Text Structuring
3. 句子聚合 - Sentence Aggregation
4. 语法化 - Lexicalisation
5. 参考表达式生成 - Referring Expression Generation | REG
6. 语言实现 - Linguistic Realisation

想要深入了解NLG，可以看看这篇文章《[一文看懂自然语言生成 - NLG（6个实现步骤+3个典型应用）](#)》

NLP 的5个难点

NLP 的5个难点

- 1 没有规律
- 2 自由组合
- 3 开放集合
- 4 知识依赖
- 5 上下文

easyai

1. 语言是没有规律的，或者说规律是错综复杂的。
2. 语言是可以自由组合的，可以组合复杂的语言表达。
3. 语言是一个开放集合，我们可以任意的发明创造一些新的表达方式。
4. 语言需要联系到实践知识，有一定的知识依赖。
5. 语言的使用要基于环境和上下文。

NLP 的4个典型应用



情感分析



聊天机器人



语音识别



机器翻译

情感分析

互联网上有大量的文本信息，这些信息想要表达的内容是五花八门的，但是他们抒发的情感是一致的：正面/积极的 - 负面/消极的。

通过情感分析，可以快速了解用户的舆情情况。

聊天机器人

过去只有 Siri、小冰这些机器人，大家使用的动力并不强，只是当做一个娱乐的方式。但是最近几年智能音箱的快速发展让大家感受到了聊天机器人的价值。

而且未来随着智能家居，智能汽车的发展，聊天机器人会有更大的使用价值。

语音识别

语音识别已经成为了全民级的引用，微信里可以语音转文字，汽车中使用导航可以直接说目的地，老年人使用输入法也可以直接语音而不用学习拼音...

机器翻译

目前的机器翻译准确率已经很高了，大家使用 Google 翻译完全可以看懂文章的大意。传统的人肉翻译未来很可能会失业。

NLP 的 2 种途径、3 个核心步骤

NLP 可以使用传统的机器学习方法来处理，也可以使用深度学习的方法来处理。2 种不同的途径也对应着不同的处理步骤。详情如下：

方式 1：传统机器学习的 NLP 流程

传统机器学习的 NLP 流程



easyai

1. 语料预处理

1. 中文语料预处理 4 个步骤（下文详解）
2. 英文语料预处理的 6 个步骤（下文详解）

2. 特征工程

1. 特征提取
2. 特征选择

3. 选择分类器

方式 2：深度学习的 NLP 流程

深度学习的 NLP 流程



easyai

1. 语料预处理

1. 中文语料预处理 4 个步骤（下文详解）
2. 英文语料预处理的 6 个步骤（下文详解）

2. 设计模型

3. 模型训练

英文 NLP 语料预处理的 6 个步骤

英文语料预处理的6个核心步骤

- ① 分词
- ② 词干提取
- ③ 词形还原
- ④ 词性标注
- ⑤ 命名实体识别
- ⑥ 分块

easyai

1. 分词 - Tokenization
2. 词干提取 - Stemming
3. 词形还原 - Lemmatization
4. 词性标注 - Parts of Speech
5. 命名实体识别 - NER
6. 分块 - Chunking

中文 NLP 语料预处理的 4 个步骤

中文语料预处理的 4 个核心步骤

- ① 分词
- ② 词性标注
- ③ 命名实体识别
- ④ 去除停用词

easyai

1. 中文分词 - Chinese Word Segmentation

2. 词性标注 - Parts of Speech
3. 命名实体识别 - NER
4. 去除停用词

总结

自然语言处理（NLP）就是在机器语言和人类语言之间沟通的桥梁，以实现人机交流的目的。

NLP的2个核心任务：

1. 自然语言理解 - NLU
2. 自然语言生成 - NLG

NLP 的5个难点：

1. 语言是没有规律的，或者说规律是错综复杂的。
2. 语言是可以自由组合的，可以组合复杂的语言表达。
3. 语言是一个开放集合，我们可以任意的发明创造一些新的表达方式。
4. 语言需要联系到实践知识，有一定的知识依赖。
5. 语言的使用要基于环境和上下文。

NLP 的4个典型应用：

1. 情感分析
2. 聊天机器人
3. 语音识别
4. 机器翻译

NLP 的6个实现步骤：

1. 分词-tokenization
2. 词干提取-stemming

3. 词形还原-lemmatization
4. 词性标注-pos tags
5. 命名实体识别-ner
6. 分块-chunking



easyAI – 产品经理的 AI 知识库
<https://easyai.tech>



公众号: easyai-tech

自然语言理解 - NLU | NLI



一文看懂自然语言理解-NLU

概念 – 应用 – 难点 – 实现方式

easyai

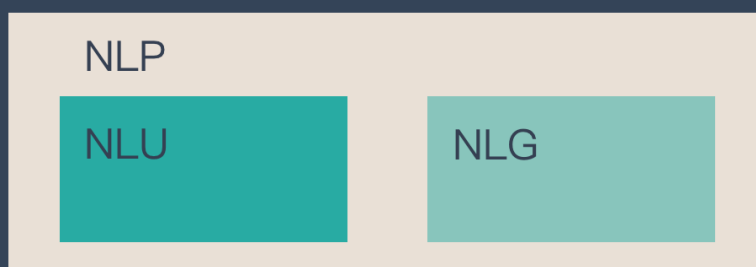
自然语言理解(NLU)跟 NLP 是什么关系？为什么说它是人工智能领域里一个难点？
NLU 的发展史历史和目前最现金的方法是什么？

本文将解答上面的问题，带你全面了解自然语言理解(NLU)。

什么是自然语言理解(NLU)?

大家最常听到的是 NLP，而 自然语言理解（NLU）则是 NLP 的一部分：

NLU 可以理解为 NLP 的一部分



easyai

什么是自然语言？

自然语言就是大家平时在生活中常用的表达方式，大家平时说的「讲人话」就是这个意思。

自然语言：我背有点驼(非自然语言：我的背部呈弯曲状)

自然语言：宝宝的经纪人睡了宝宝的宝宝

自然语言理解就是希望机器像人一样，具备正常人的语言理解能力，由于自然语言在理解上有很多难点(下面详细说明)，所以 NLU 是至今还远不如人类的表现。



自然语言理解就是希望机器像人一样，具备正常人的语言理解能力

easyai

下面用一个具体的案例来深度说明一下自然语言理解（NLU）：

对话系统这个事情在2015年开始突然火起来了，主要是因为一个技术的普及：**机器学习特别是深度学习带来的语音识别和NLU(自然语言理解)**——主要解决的是识别人讲的话。

这个技术的普及让很多团队都掌握了一组关键技能：**意图识别和实体提取。**

这意味着什么？我们来看一个例子。

在生活中，如果想要订机票，人们会有很多种自然的表达：

“订机票”；

“有去上海的航班么？”；

“看看航班，下周二出发去纽约的”；

“要出差，帮我查下机票”；

等等等等

可以说“自然的表达”有无穷多的组合（自然语言）都是在代表“订机票”这个意图的。而听到这些表达的人，可以准确理解这些表达指的是“订机票”这件事。

而要**理解这么多不同的表达，对机器是个挑战。**在过去，机器只能处理“结构化的数据”（比如关键词），也就是说如果要听懂人在讲什么，必须要用户输入精确的指令。

所以，无论你说“我要出差”还是“帮我看看去北京的航班”，只要这些字里面没有包含提前设定好的关键词“订机票”，系统都无法处理。而且，只要出现了关键词，比如“我要退订机票”里也有这三个字，也会被处理成用户想要订机票。

基于规则的意图判断

机器通过「订机票」这个关键词来识别意图

我想订机票



我要出差



帮我看看去北京的航班



自然语言理解这个技能出现后，可以让机器从各种自然语言的表达中，区分出来，哪些话归属于这个意图；而那些表达不是归于这一类的，而不再依赖那么死板的关键词。比如经过训练后，机器能够识别“帮我推荐一家附近的餐厅”，就不属于“订机票”这个意图的表达。

并且，通过训练，机器还能够在句子当中自动提取出来“上海”，这两个字指的是目的地这个概念（即实体）；“下周二”指的是出发时间。

这样一来，看上去“机器就能听懂人话啦！”。

基于 NLU 来识别用户意图

“看看航班，下周二出发去纽约的

意图：订机票
时间：下周二
目的地：纽约

easyai

自然语言理解（NLU）的应用

几乎所有跟文字语言和语音相关的应用都会用到 NLU，下面举一些具体的例子。

基于 NLU 的应用



机器翻译



机器客服



智能音箱

easyai

机器翻译

基于规则的翻译效果经常不太好，所以如果想提升翻译的效果，必须建立在对内容的理解之上。

如果是不理解上下文，就会出现下面的笑话：

I like apple, it's so fast!

我喜欢「苹果」，它很快！

机器客服

如果想实现问答，就要建立在多轮对话的理解基础之上，自然语言理解是必备的能力。

下面的例子对于机器来说就很难理解：

"有什么可以帮您？"

"你好，我想投诉"

"请问投诉的车牌号是多少？"

"XXXXXX"

"请问是什么问题？"

"我刚上车，那个态度恶劣的哥谭市民就冲我发火"

机器很容易理解为：那个态度恶劣/的/哥谭/市民/就冲我发火

智能音箱

智能音箱中，NLU 也是重要的一个环节。很多语音交互都是很短的短语，音箱不但需要能否识别用户在说什么话，更要理解用户的意图。

"我冷了"

用户并没有提到空调，但是机器需要知道用户的意图——空调有点冷，需要把温度调高。

自然语言理解（NLU）的难点

下面先列举一些机器不容易理解的案例：

1. 校长说衣服上除了校徽别别别的
2. 过几天天天天气不好
3. 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。
4. 今天多得谢逊出手相救，在这里我想真心感谢“谢谢谢逊大侠出手”
5. 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
6. 姑姑你估估我鼓鼓的口袋里有多少谷和菇！！
7. “你看到王刚了吗”“王刚刚刚刚走”
8. 张杰陪俩女儿跳格子：俏俏我们不要跳跳跳跳过的格子啦

自然语言理解的5大难点

1. 语言的多样性
2. 语言的歧义性
3. 语言的鲁棒性
4. 语言的知识依赖
5. 语言的上下文

那么对于机器来说，NLU 难点大致可以归为5类：

难点1：语言的多样性

自然语言没有什么通用的规律，你总能找到很多例外的情况。

另外，自然语言的组合方式非常灵活，字、词、短语、句子、段落...不同的组合可以表达出很多的含义。例如：

我要听大王叫我来巡山

给我播大王叫我来巡山

我想听歌大王叫我来巡山

放首大王叫我来巡山

给唱一首大王叫我来巡山

放音乐大王叫我来巡山

放首歌大王叫我来巡山

给大爷来首大王叫我来巡山

难点2：语言的歧义性

如果不联系上下文，缺少环境的约束，语言有很大的歧义性。例如：

我要去拉萨

- 需要火车票？
- 需要飞机票？
- 想听音乐？
- 还是想查找景点？

难点3：语言的鲁棒性

自然语言在输入的过程中，尤其是通过语音识别获得的文本，会存在多字、少字、错字、噪音等问题。例如：

大王叫我来新山

大王叫让我来巡山

大王叫我巡山

难点4：语言的知识依赖

语言是对世界的符号化描述，语言天然连接着世界知识，例如：

大鸭梨

除了表示水果，还可以表示餐厅名

7天

可以表示时间，也可以表示酒店名

晚安

有一首歌也叫《晚安》

难点5：语言的上下文

上下文的概念包括很多种：对话的上下文、设备的上下文、应用的上下文、用户画像...

U：买张火车票

A：请问你要去哪里？

U：宁夏

U：来首歌听

A：请问你想听什么歌？

U：宁夏

NLU 的实现方式

自然语言理解跟整个人工智能的发展历史类似，一共经历了3次迭代：

1. 基于规则的方法
2. 基于统计的方法
3. 基于深度学习的方法

自然语言理解3大发展阶段



阶段1：基于规则



阶段2：基于统计



阶段3：基于深度学习

easyai

最早大家通过总结规律来判断自然语言的意图，常见的方法有：CFG、JSGF等。

后来出现了基于统计学的 NLU 方式，常见的方法有：SVM、ME等。

随着深度学习的爆发，CNN、RNN、LSTM 都成为了最新的“统治者”。

到了2019年，BERT 和 GPT-2 的表现震惊了业界，他们都是用了 Transformer，下面将重点介绍 Transformer，因为他是目前「最先进」的方法。



截止2019年，Transformer 是当红炸子鸡

easyai

Transformer 和 CNN / RNN 的比较

Transformer 的原理比较复杂，这里就不详细说明了，感兴趣的朋友可以查看下面的文章，讲的很详细：

《BERT大火却不懂Transformer? 读这一篇就够了》

下面将摘取一部分《why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures》里的数据，直观的让大家看出来3者的比较。

语义特征提取能力

语义特征抽取能力：Transformer>>原生CNN=原生RNN

原生RNN

原生CNN

Model	DE→EN				DE→FR			
	PPL	2014	2017	Acc(%)	PPL	2012	Acc(%)	
<i>RNNS2S</i>	5.7	29.1	30.1	84.0	7.06	16.4	72.2	
<i>ConvS2S</i>	6.3	29.1	30.4	82.3	7.93	16.8	72.7	
<i>Transformer</i>	4.3	32.7	33.7	90.3	4.9	18.7	76.7	
<i>uedin-wmt17</i>	—	—	35.1	87.9	—	—	—	
<i>TransRNN</i>	5.2	30.5	31.9	86.1	6.3	17.6	74.2	

Table 5: The results of different architectures on *newstest* sets and *ContraWSD*. *PPL* is the perplexity on the validation set. *Acc* means accuracy on the test set.

easyai

从语义特征提取能力来说，目前实验支持如下结论：Transformer在这方面的能力非常显著地超过RNN和CNN（在考察语义类能力的任务WSD中，Transformer超过RNN和CNN大约4-8个绝对百分点），RNN和CNN两者能力差不太多。

长距离特征捕获能力

长距离特征抽取能力：Transformer>原生RNN>原生CNN

原生RNN

原生CNN

Model	2014	2017	PPL	Acc(%)
<i>RNNS2S</i>	23.3	25.1	6.1	95.1
<i>ConvS2S</i>	23.9	25.2	7.0	84.9
<i>Transformer</i>	26.7	27.5	4.5	97.1
<i>RNN-bideep</i>	24.7	26.1	5.7	96.3

Table 2: The results of different NMT models, including the BLEU scores on *newstest2014* and *newstest2017*, the perplexity on the validation set, and the accuracy of long-range dependencies.

easyai

原生CNN特征抽取器在这方面极为显著地弱于RNN和Transformer，Transformer微弱优于RNN模型（尤其在主语谓语距离小于13时），能力由强到弱排序为Transformer>RNN>>CNN; 但在比较远的距离上（主语谓语距离大于13），RNN微弱优于Transformer，所以综合看，可以认为Transformer和RNN在这方面能力差不太多，而CNN则显著弱于前两者。

任务综合特征抽取能力

任务综合特征抽取能力：Transformer>>原生CNN=原生RNN

原生RNN

原生CNN

Model	DE→EN				DE→FR		
	PPL	2014	2017	Acc(%)	PPL	2012	Acc(%)
RNNS2S	5.7	29.1	30.1	84.0	7.06	16.4	72.2
ConvS2S	6.3	29.1	30.4	82.3	7.93	16.8	72.7
Transformer	4.3	32.7	33.7	90.3	4.9	18.7	76.7
uedin-wmt17	—	—	35.1	87.9	—	—	—
TransRNN	5.2	30.5	31.9	86.1	6.3	17.6	74.2

Table 5: The results of different architectures on *newstest* sets and *ContraWSD*. *PPL* is the perplexity on the validation set. *Acc* means accuracy on the test set.

easyai

Transformer综合能力要明显强于RNN和CNN（你要知道，技术发展到现在阶段，BLEU绝对值提升1个点是很难的事情），而RNN和CNN看上去表现基本相当，貌似CNN表现略好一些。

并行计算能力及运算效率

计算效率：Transformer>CNN>RNN

原生RNN

原生CNN

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(\log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

easyai

Transformer Base最快，CNN次之，再次Transformer Big，最慢的是RNN。RNN比前两者慢了3倍到几十倍之间。

关于 Transformer，推荐几篇优秀的文章给大家，让大家有一个更综合的了解：

《放弃幻想，全面拥抱Transformer：自然语言处理三大特征抽取器（CNN/RNN/TF）比较》

《从Word Embedding到Bert模型—自然语言处理中的预训练技术发展史》

《效果惊人的GPT 2.0模型：它告诉了我们什么》



easyAI – 产品经理的 AI 知识库
<https://easyai.tech>



公众号：easyai-tech

自然语言生成 - NLG



一文看懂自然语言生成 – NLG

基本概念

6个实现步骤

3个典型应用

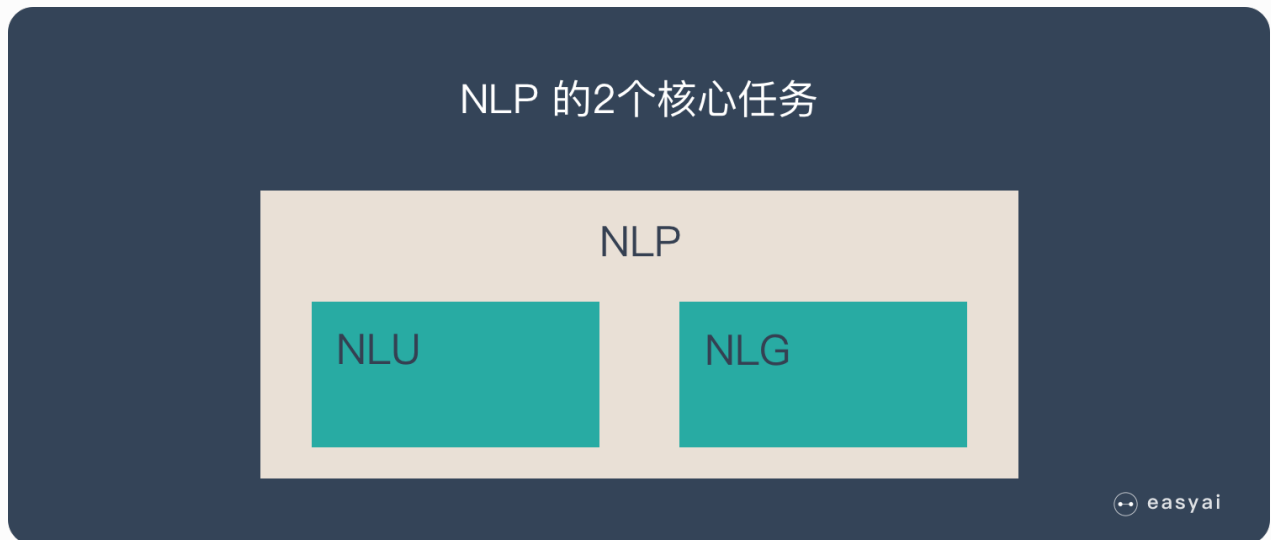
easyai

自然语言生成 - NLG 是 NLP 的重要组成部分，他的主要目的是降低人类和机器之间的沟通鸿沟，将非语言格式的数据转换成人类可以理解的语言格式。

本文除了介绍 NLG 的基本概念，还会介绍 NLG 的3个 Level、6个步骤和3个典型的应用。

什么是 NLG?

NLG 是 NLP 的一部分



NLP = NLU + NLG

自然语言生成 - NLG 是 NLP 的重要组成部分。NLU 负责理解内容，NLG 负责生成内容。

以智能音箱为例，当用户说“几点了？”，首先需要利用 NLU 技术判断用户意图，理解用户想要什么，然后利用 NLG 技术说出“现在是6点50分”。

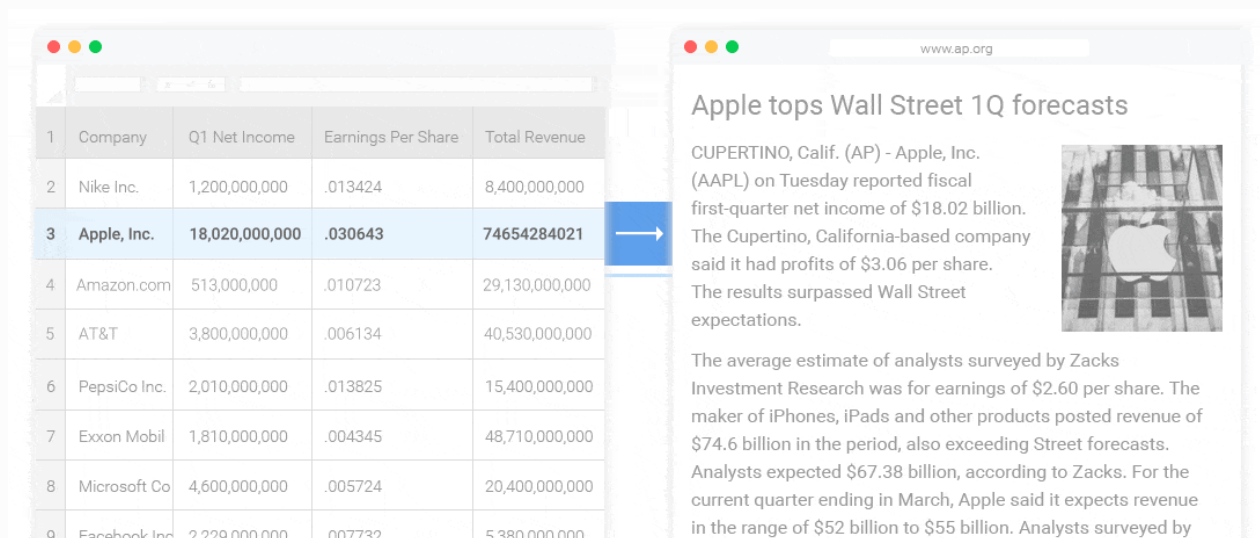
自然语言生成 - NLG 是什么？



NLG 是为了跨越人类和机器之间的沟通鸿沟，将非语言格式的数据转换成人类可以理解的语言格式，如文章、报告等。

自然语言生成 - NLG 有2种方式：

1. text - to - text：文本到语言的生成
2. data - to - text：数据到语言的生成



	Company	Q1 Net Income	Earnings Per Share	Total Revenue
1				
2	Nike Inc.	1,200,000,000	.013424	8,400,000,000
3	Apple, Inc.	18,020,000,000	.030643	74654284021
4	Amazon.com	513,000,000	.010723	29,130,000,000
5	AT&T	3,800,000,000	.006134	40,530,000,000
6	PepsiCo Inc.	2,010,000,000	.013825	15,400,000,000
7	Exxon Mobil	1,810,000,000	.004345	48,710,000,000
8	Microsoft Co	4,600,000,000	.005724	20,400,000,000
9	Facebook Inc	2,229,000,000	.007732	5,380,000,000

Apple tops Wall Street 1Q forecasts

CUPERTINO, Calif. (AP) - Apple, Inc. (AAPL) on Tuesday reported fiscal first-quarter net income of \$18.02 billion. The Cupertino, California-based company said it had profits of \$3.06 per share. The results surpassed Wall Street expectations.

The average estimate of analysts surveyed by Zacks Investment Research was for earnings of \$2.60 per share. The maker of iPhones, iPads and other products posted revenue of \$74.6 billion in the period, also exceeding Street forecasts. Analysts expected \$67.38 billion, according to Zacks. For the current quarter ending in March, Apple said it expects revenue in the range of \$52 billion to \$55 billion. Analysts surveyed by

NLG 的3个 LEVEL



简单的数据合并：自然语言处理的简化形式，这将允许将数据转换为文本（通过类似 Excel 的函数）。为了关联，以**邮件合并**（MS Word mailmerge）为例，其中间隙填充了一些数据，这些数据是从另一个源（例如 MS Excel 中的表格）中检索的。

文档2 - Word

文件 开始 插入 设计 布局 引用 邮件 审阅 视图 办公标签 ABBYY FineReader 12 设计 布局 告诉我想要做什么 共享

中文信封 信封 标签 开始 选择 编辑 突出显示 地址块 问候语 插入 规则 匹配域 预览结果 查找收件人 完成并合并 检查错误

创建 邮件合并 - 收件人 - 收件人列表 开始邮件合并 编写和插入域 预览结果 完成

文档2 x

CEACA 国家信息化计算机教育认证

准考证

个人信息	
姓 名	I
性 别	
身份证号	
专业信息	
班 级	
专 业	
考试信息	
考试模块	
准考证号	
登录口令	
考 场	
考试时间	

节: 1 第 1 页, 共 1 页 58 个字 中文(中国) 120%

模板化的 NLG：这种形式的NLG使用模板驱动模式来显示输出。以足球比赛得分板为例。数据动态地保持更改，并由预定义的业务规则集（如if / else循环语句）生成。

北京 今日 晴热高温 ， 及时添衣，谨防感冒

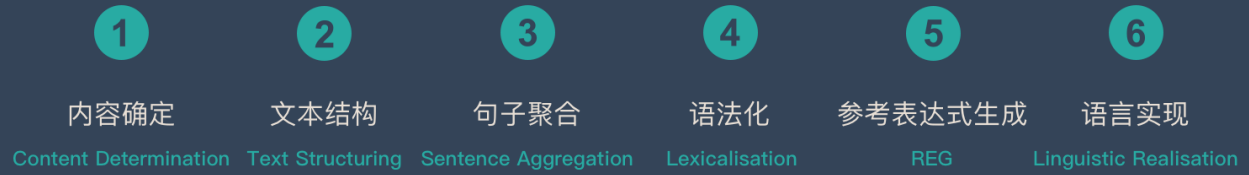
一起来看下今日天气

今天是 6月29日 ， 北京 的天气 晴热高温 ， 最高温 33 ℃，最低温是 23 ℃。 各项气象条件适宜，发生感冒机率较低。但请避免长期处于空调房间中，以防感冒 。 接下来几天， 气温保持平稳 。

高级 NLG：这种形式的自然语言生成就像人类一样。它理解意图，添加智能，考虑上下文，并将结果呈现在用户可以轻松阅读和理解的富有洞察力的叙述中。

NLG 的6个步骤

NLG 的6个步骤



easyai

第一步：内容确定 - Content Determination

作为第一步，NLG 系统需要决定哪些信息应该包含在正在构建的文本中，哪些不应该包含。通常数据中包含的信息比最终传达的信息要多。

第二步：文本结构 - Text Structuring

确定需要传达哪些信息后，NLG 系统需要合理的组织文本的顺序。例如在报道一场篮球比赛时，会优先表达「什么时间」「什么地点」「哪2支球队」，然后再表达「比赛的概况」，最后表达「比赛的结局」。

第三步：句子聚合 - Sentence Aggregation

不是每一条信息都需要一个独立的句子来表达，将多个信息合并到一个句子里表达可能会更加流畅，也更易于阅读。

第四步：语法化 - Lexicalisation

当每一句的内容确定下来后，就可以将这些信息组织成自然语言了。这个步骤会在各种信息之间加一些连接词，看起来更像是一个完整的句子。

第五步：参考表达式生成 - Referring Expression Generation | REG

这个步骤跟语法化很相似，都是选择一些单词和短语来构成一个完整的句子。不过他跟语法化的本质区别在于“REG需要识别出内容的领域，然后使用该领域（而不是其他领域）的词汇”。

第六步：语言实现 - Linguistic Realisation

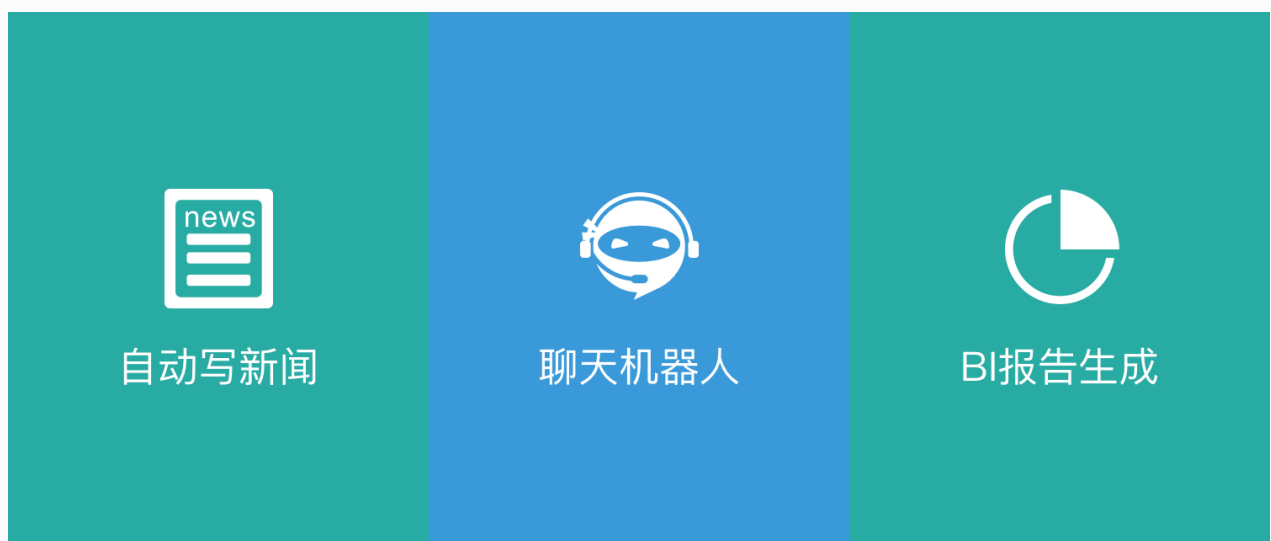
最后，当所有相关的单词和短语都已经确定时，需要将它们组合起来形成一个结构良好的完整句子。

NLG 的3种典型应用

NLG 的不管如何应用，大部分都是下面的3种目的：

1. 能够大规模的产生个性化内容
2. 帮助人类洞察数据，让数据更容易理解
3. 加速内容生产

下面给大家列一些比较典型的应用：



自动写新闻

某些领域的新闻是有比较明显的规则的，比如体育新闻。目前很多新闻已经借助 NLG 来完成了。

《腾讯机器人日均写稿过千篇 你读的新闻可能是AI写的》

聊天机器人

大家了解聊天机器人都是从 Siri 开始的，最近几年又出现了智能音箱的热潮。

除了大家日常生活中很熟悉的领域，客服工作也正在被机器人替代，甚至一些电话客服也是机器人。

《跟你通话的客服是个机器人！》



BI 的解读和报告生成

几乎各行各业都有自己的数据统计和分析工具。这些工具可以产生各式各样的图表，但是输出结论和观点还是需要依赖人。NLG 的一个很重要的应用就是解读这些数据，自动的输出结论和观点。（如下图所示）



总结

自然语言生成 - NLG 是 NLP 的重要组成部分，他的主要目的是降低人类和机器之间的沟通鸿沟，将非语言格式的数据转换成人类可以理解的语言格式。

NLG 的3个level:

1. 简单的数据合并
2. 模块化的 NLG
3. 高级 NLG

NLG 的6个步骤:

1. 内容确定 - Content Determination
2. 文本结构 - Text Structuring
3. 句子聚合 - Sentence Aggregation
4. 语法化 - Lexicalisation
5. 参考表达式生成 - Referring Expression Generation | REG

6. 语言实现 - Linguistic Realisation

NLG 应用的3个目的：

1. 能够大规模的产生个性化内容
2. 帮助人类洞察数据，让数据更容易理解
3. 加速内容生产

NLG 的3个典型应用：

1. 自动写新闻
2. 聊天机器人
3. BI 的解读和报告生成



easyAI – 产品经理的 AI 知识库
<https://easyai.tech>



公众号：easyai-tech

分词 - TOKENIZATION

easyai.tech

很棒

一文看懂中英文分词

3 大难点 中英文分词的 3 大区别 3 种典型方法

分词是 NLP 的基础任务，将句子，段落分解为字词单位，方便后续的处理的分析。

本文将介绍分词的原因，中英文分词的3个区别，中文分词的3大难点，分词的3种典型方法。最后将介绍中文分词和英文分词常用的工具。

什么是分词？

分词是 **自然语言理解 - NLP** 的重要步骤。

分词就是将句子、段落、文章这种长文本，分解为以字词为单位的数据结构，方便后续的处理分析工作。

easyai.tech 是很棒的人工智能科普网站



easyai.tech \ 是 \ 很棒的 \ 人工智能 \ 科普 \ 网站

easyai

为什么要分词？

1. 将复杂问题转化为数学问题

在 **机器学习的文章** 中讲过，机器学习之所以看上去可以解决很多复杂的问题，是因为它把这些问题都转化为了数学问题。

而 NLP 也是相同的思路，文本都是一些「非结构化数据」，我们需要先将这些数据转化为「结构化数据」，结构化数据就可以转化为数学问题了，而分词就是转化的第一步。



2. 词是一个比较合适的粒度

词是表达完整含义的最小单位。

字的粒度太小，无法表达完整含义，比如“鼠”可以是“老鼠”，也可以是“鼠标”。

而句子的粒度太大，承载的信息量多，很难复用。比如“传统方法要分词，一个重要原因是传统方法对远距离依赖的建模能力较弱。”



3. 深度学习时代，部分任务中也可以「分字」

深度学习时代，随着数据量和算力的爆炸式增长，很多传统的方法被颠覆。

分词一直是 NLP 的基础，但是现在也不一定了，感兴趣的可以看看这篇论文：《[Is Word Segmentation Necessary for Deep Learning of Chinese Representations?](#)》。

Is Word Segmentation Necessary for Deep Learning of Chinese Representations?

Yuxian Meng^{*1}, Xiaoya Li^{*1}, Xiaofei Sun¹, Qinghong Han¹
Arianna Yuan^{1,2}, and Jiwei Li¹

¹ Shannon.AI

² Computer Science Department, Stanford University

{ yuxian_meng, xiaoya_li, xiaofei_sun, qinghong_han
arianna_yuan, jiwei_li }@shannonai.com

不过在一些特定任务中，分词还是必要的。如：关键词提取、命名实体识别等。

中英文分词的3个典型区别

中英文分词的 3 大区别



分词方式不同，「中文」更难



「英文」单词有多重形态



「中文」需要考虑分词粒度

easyai

区别1：分词方式不同，中文更难

英文有天然的空格作为分隔符，但是中文没有。所以如何切分是一个难点，再加上中文里一词多意的情况非常多，导致很容易出现歧义。下文中难点部分会详细说明。

区别2：英文单词有多种形态

英文单词存在丰富的变形变换。为了应对这些复杂的变换，英文NLP相比中文存在一些独特的处理步骤，我们称为词形还原（Lemmatization）和词干提取（Stemming）。中文则不需要

词性还原：does, done, doing, did 需要通过词性还原恢复成 do。

词干提取：cities, children, teeth 这些词，需要转换为 city, child, tooth”这些基本形态

区别3：中文分词需要考虑粒度问题

例如「中国科学技术大学」就有很多种分法：

- 中国科学技术大学
- 中国 \ 科学技术 \ 大学
- 中国 \ 科学 \ 技术 \ 大学

粒度越大，表达的意思就越准确，但是也会导致召回比较少。所以中文需要不同的场景和要求选择不同的粒度。这个在英文中是没有的。

中文分词的3大难点



难点 1：没有统一的标准

目前中文分词没有统一的标准，也没有公认规范。不同的公司和组织各有各的方法和规则。

难点 2：歧义词如何切分

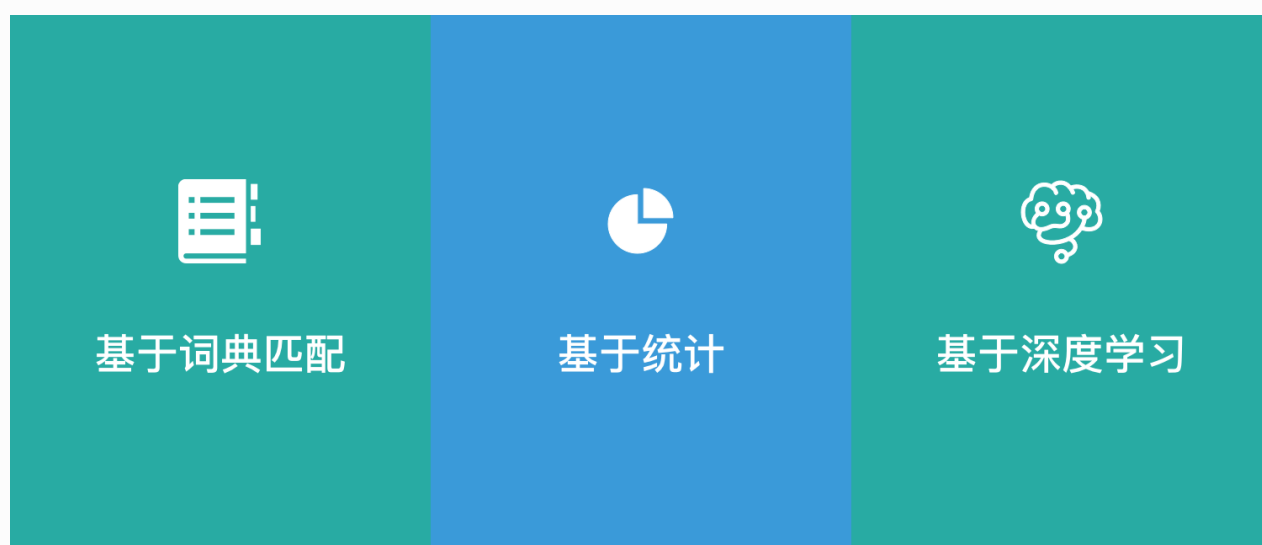
例如「兵乓球拍卖完了」就有2种分词方式表达了2种不同的含义：

- 乒乓球 \ 拍卖 \ 完了
- 乒乓 \ 球拍 \ 卖 \ 完了

难点 3：新词的识别

信息爆炸的时代，三天两头就会冒出来一堆新词，如何快速的识别出这些新词是一大难点。比如当年「蓝瘦香菇」大火，就需要快速识别。

3种典型的分词方法



分词的方法大致分为 3 类：

1. 基于词典匹配
2. 基于统计
3. 基于深度学习

给予词典匹配的分词方式

优点：速度快、成本低

缺点：适应性不强，不同领域效果差异大

基本思想是基于词典匹配，将待分词的中文文本根据一定规则切分和调整，然后跟词典中的词语进行匹配，匹配成功则按照词典的词分词，匹配失败通过调整或者重新选择，如此反复循环即可。代表方法有基于正向最大匹配和基于逆向最大匹配及双向匹配法。

基于统计的分词方法

优点：适应性较强

缺点：成本较高，速度较慢

这类目前常用的是算法是**HMM**、**CRF**、**SVM**、**深度学习**等算法，比如stanford、Hanlp分词工具是基于CRF算法。以CRF为例，基本思路是对汉字进行标注训练，不仅考虑了词语出现的频率，还考虑上下文，具备较好的学习能力，因此其对歧义词和未登录词的识别都具有良好的效果。

基于深度学习

优点：准确率高、适应性强

缺点：成本高，速度慢

例如有人员尝试使用双向LSTM+CRF实现分词器，其本质上是序列标注，所以有通用性，命名实体识别等都可以使用该模型，据报道其分词器字符准确率可高达97.5%。

常见的分词器都是使用机器学习算法和词典相结合，一方面能够提高分词准确率，另一方面能够改善领域适应性。

中文分词工具

下面排名根据 GitHub 上的 star 数排名：

1. Hanlp
2. Stanford 分词
3. ansj 分词器
4. 哈工大 LTP
5. KCWS分词器
6. jieba
7. IK
8. 清华大学THULAC
9. ICTCLAS

英文分词工具

1. Keras
2. Spacy
3. Gensim
4. NLTK

总结

分词就是将句子、段落、文章这种长文本，分解为以字词为单位的数据结构，方便后续的处理分析工作。

分词的原因：

1. 将复杂问题转化为数学问题
2. 词是一个比较合适的粒度
3. 深度学习时代，部分任务中也可以「分字」

中英文分词的3个典型区别：

1. 分词方式不同，中文更难
2. 英文单词有多种形态，需要词性还原和词干提取
3. 中文分词需要考虑粒度问题

中文分词的3大难点

1. 没有统一的标准
2. 歧义词如何切分
3. 新词的识别

3个典型的分词方式：

1. 基于词典匹配
2. 基于统计
3. 基于深度学习



easyAI – 产品经理的 AI 知识库
<https://easyai.tech>



公众号：easyai-tech

词干提取STEMMING和词形还原 LEMMATISATION



一文看懂 词干提取+词形还原

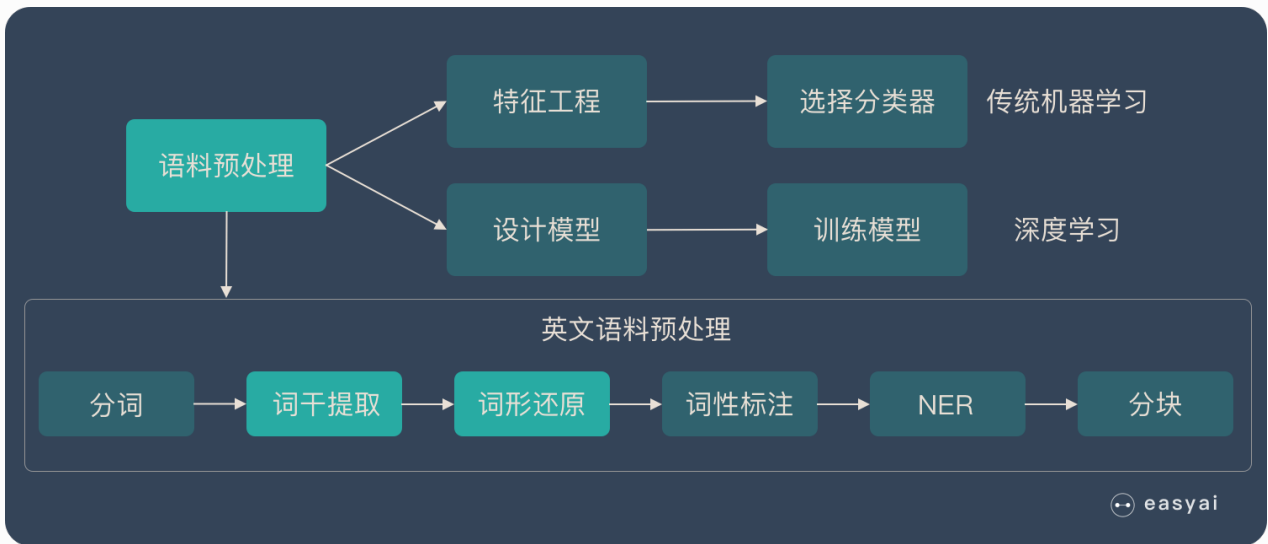
↔ easyai

词干提取和词形还原是英文语料预处理中的重要环节。虽然他们的目的一致，但是两者还是存在一些差异。

本文将介绍他们的概念、异同、实现算法等。

词干提取和词形还原在 NLP 中在什么位置？

词干提取是英文语料预处理的一个步骤（中文并不需要），而语料预处理是 NLP 的第一步，下面这张图将让大家知道词干提取在这个知识结构中的位置。



什么是词干提取和词形还原？

词干提取 - Stemming

词干提取是去除单词的前后缀得到词根的过程。

大家常见的前后词缀有「名词的复数」、「进行式」、「过去分词」...

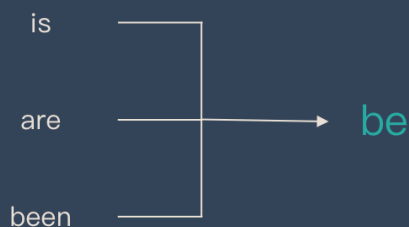


词形还原 - Lemmatisation

词形还原是基于词典，将单词的复杂形态转变成最基础的形态。

词形还原不是简单地将前后缀去掉，而是会根据词典将单词进行转换。比如「drove」会转换为「drive」。

词形还原 – Lemmatisation



easyai

为什么要做词干提取和词形还原？

比如当我搜索「play basketball」时，Bob is playing basketball 也符合我的要求，，但是 play 和 playing 对于计算机来说是 2 种完全不同的东西，所以我们需要将 playing 转换成 play。

词干提取和词形还原的目的就是将长相不同，但是含义相同的词统一起来，这样方便后续的处理和分析。

词干提取和词形还原的 4 个相似点



1. 目标一致。词干提取和词形还原的目标均为将词的屈折形态或派生形态简化或归并为词干（stem）或原形的基础形式，都是一种对词的不同形态的统一归并的过程。
2. 结果部分交叉。词干提取和词形还原不是互斥关系，其结果是有部分交叉的。一部分词利用这两类方法都能达到相同的词形转换效果。如“dogs”的词干为“dog”，其原形也为“dog”。
3. 主流实现方法类似。目前实现词干提取和词形还原的主流实现方法均是利用语言中存在的规则或利用词典映射提取词干或获得词的原形。
4. 应用领域相似。主要应用于信息检索和文本、自然语言处理等方面，二者均是这些应用的基本步骤。

词干提取和词形还原的 5 个不同点



1. 在原理上，词干提取主要是采用“缩减”的方法，将词转换为词干，如将“cats”处理为“cat”，将“effective”处理为“effect”。而词形还原主要采用“转变”的方法，将词转变为其原形，如将“drove”处理为“drive”，将“driving”处理为“drive”。
2. 在复杂性上，词干提取方法相对简单，词形还原则需要返回词的原形，需要对词形

进行分析，不仅要进行词缀的转化，还要进行词性识别，区分相同词形但原形不同的词的差别。词性标注的准确率也直接影响词形还原的准确率，因此，词形还原更为复杂。

3. 在实现方法上，虽然词干提取和词形还原实现的主流方法类似，但二者在具体实现上各有侧重。词干提取的实现方法主要利用规则变化进行词缀的去除和缩减，从而达到词的简化效果。词形还原则相对较复杂，有复杂的形态变化，单纯依据规则无法很好地完成。其更依赖于词典，进行词形变化和原形的映射，生成词典中的有效词。
4. 在结果上，词干提取和词形还原也有部分区别。词干提取的结果可能并不是完整的、具有意义的词，而只是词的一部分，如“revival”词干提取的结果为“reviv”，“airliner”词干提取的结果为“airlin”。而经词形还原处理后获得的结果是具有一定意义的、完整的词，一般为词典中的有效词。
5. 在应用领域上，同样各有侧重。虽然二者均被应用于信息检索和文本处理中，但侧重不同。词干提取更多被应用于信息检索领域，如Solr、Lucene等，用于扩展检索，粒度较粗。词形还原更主要被应用于文本挖掘、自然语言处理，用于更细粒度、更为准确的文本分析和表达

3 种主流的词干提取算法

3 种主流的词干提取算法

Porter

 Snowball

Lancaster

 easyai

Porter

这种词干算法比较旧。它是从20世纪80年代开始的，其主要关注点是删除单词的共同结尾，以便将它们解析为通用形式。它不是太复杂，它的开发停止了。

通常情况下，它是一个很好的起始基本词干分析器，但并不建议将它用于复杂的应用。相反，它在研究中作为一种很好的基本词干算法，可以保证重复性。与其他算法相比，它也是一种非常温和的词干算法。

「推荐」Snowball

种算法也称为 Porter2 词干算法。它几乎被普遍认为比 Porter 更好，甚至发明 Porter 的开发者也这么认为。Snowball 在 Porter 的基础上加了很多优化。Snowball 与 Porter 相比差异约为5%。

Lancaster

Lancaster 的算法比较激进，有时候会处理成一些比较奇怪的单词。如果在 NLTK 中使用词干分析器，则可以非常轻松地将自己的自定义规则添加到此算法中。

词形还原的实践方法

词形还原是基于词典的，每种语言都需要经过语义分析、词性标注来建立完整的词库，目前英文词库是很完善的。

Python 中的 NLTK 库包含英语单词的词汇数据库。这些单词基于它们的语义关系链接在一起。链接取决于单词的含义。特别是，我们可以利用 WordNet。

```
import nltk
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
print(lemmatizer.lemmatize("blogs"))
#Returns blog
```

总结

词干提取和词形还原都是将长相不同，但是含义相同的词统一起来，这样方便后续的处理和分析。

他们是英文语料预处理中的一个环节。

词干提取和词形还原的 4 个相似点：

1. 目标一致
2. 部分结果一致

3. 主流实现方式类似
4. 应用领域相似

词干提取和词形还原的 5 个不同点：

1. 原理上不同
2. 词形还原更加复杂
3. 具体实现方式的侧重点不同
4. 呈现结果有区别
5. 应用领域上，侧重点不完全一致

3 种词干提取的主流算法：

1. Porter
2. Snowball
3. Lancaster

英文的词形还原可以直接使用 Python 中的 NLTK 库，它包含英语单词的词汇数据库。



easyAI – 产品经理的 AI 知识库
<https://easyai.tech>



公众号：easyai-tech

词性标注 – PART OF SPEECH

分词 及 词性

标注

■ 名词 ■ 连词
■ 动词

一文看懂词性标注

Part of speech

easyai

什么是词性标注？

分词词性											
谢尔盖·科罗廖夫 PER	(w	1907年1月12日 TIME	- w	1966年1月14日 TIME) w	, w	原 a	苏联 LOC	宇航 n		
事业 n	的 u	伟大 a	设计师 n	与 c	组织者 n	, w	第一枚 m	射程 n	超过 v	8000公里 m	
的 u	洲际 a	火箭 n	(w	弹道导弹 nz) w	的 u	设计者 n	, w	第一颗 m	人造地球卫星 nz	
的 u	运载火箭 n	的 u	设计者 n	、 w	第一艘 m	载人航天 vn	飞船 n	的 u	总设计师 n	。 w	

维基百科上对词性的定义为：In traditional grammar, a part of speech (abbreviated form: PoS or POS) is a category of words (or, more generally, of lexical items) which have similar grammatical properties.

词性指以词的特点作为划分词类的根据。词类是一个语言学术语，是一种语言中词的语法分类，是以语法特征（包括句法功能和形态变化）为主要依据、兼顾词汇意义对词进行划分的结果。

从组合和聚合关系来说，一个词类是指：在一个语言中，众多具有相同句法功能、能在同样的组合位置中出现的词，聚合在一起形成的范畴。词类是最普遍的语法的聚合。词类划分具有层次性。如汉语中，词可以分成实词和虚词，实词中又包括体词、谓词等，体词中又可以分出名词和代词等。

词性标注就是在给定句子中判定每个词的语法范畴，确定其词性并加以标注的过程，这也是自然语言处理中一项非常重要的基础性工作，所有对于词性标注的研究已经有较长的时间，在研究者长期的研究总结中，发现汉语词性标注中面临了许多棘手的问题。

中文词性标注的难点

汉语是一种缺乏词形态变化的语言，词的类别不能像印欧语那样，直接从词的形态变化上来判别。

常用词兼类现象严重。《现代汉语八百词》收取的常用词中，兼类词所占的比例高达22.5%，而且发现越是常用的词，不同的用法越多。由于兼类使用程度高，兼类现象涉及汉语中大部分词类，因而造成在汉语文本中词类歧义排除的任务量巨大。

研究者主观原因造成的困难。语言学界在词性划分的目的、标准等问题上还存在分歧。目前还没有一个统的被广泛认可汉语词类划分标准，词类划分的粒度和标记符号都不统一。词类划分标准和标记符号集的差异，以及分词规范的含混性，给中文信息处理带来了极大的困难。

词性标注4种常见方法



关于词性标注的研究比较多，这里介绍一波常见的几类方法，包括基于规则的词性标注方法、基于统计模型的词性标注方法、基于统计方法与规则方法相结合的词性标注方法、基于深度学习的词性标注方法等。

基于规则的词性标注方法

基于规则的词性标注方法是人们提出较早的一种词性标注方法，其基本思想是按兼类词搭配关系和上下文语境建造词类消歧规则。早期的词类标注规则一般由人工构建。

随着标注语料库规模的增大，可利用的资源也变得越来越，这时候以人工提取规则的方法显然变得不现实，于是乎，人们提出了基于机器学习的规则自动提出方法。

基于统计模型的词性标注方法

统计方法将词性标注看作是一个序列标注问题。其基本思想是：给定带有各自标注的词的序列，我们可以确定下一个词最可能的词性。

现在已经有隐马尔可夫模型（HMM）、条件随机域（CRF）等统计模型了，这些模型可以使用有标记数据的大型语料库进行训练，而有标记的数据则是指其中每一个词都分配了正确的词性标注的文本。

基于统计方法与规则方法相结合的词性标注方法

理性主义方法与经验主义相结合的处理策略一直是自然语言处理领域的专家们不断研究和探索的问题，对于词性标注问题当然也不例外。

这类方法的主要特点在于对统计标注结果的筛选，只对那些被认为可疑的标注结果，才采用规则方法进行歧义消解，而不是对所有情况都既使用统计方法又使用规则方法。

基于深度学习的词性标注方法

可以当作序列标注的任务来做，目前深度学习解决序列标注任务常用方法包括 [LSTM+CRF](#)、[BiLSTM+CRF](#)等。

值得一提的是，这一类方法近年来文章非常多，想深入了解这一块的朋友们可以看这里：[NLP-progress - GitHub](#)

最后再放一个词性标注任务数据集 – [人民日报1998词性标注数据集](#)

词性标注工具推荐

Jieba

“结巴”中文分词：做最好的 Python 中文分词组件，可以进行词性标注。

[Github地址](#)

SnowNLP

SnowNLP是一个python写的类库，可以方便的处理中文文本内容。

[Github地址](#)

THULAC

THULAC (THU Lexical Analyzer for Chinese) 由清华大学自然语言处理与社会人文计算实验室研制推出的一套中文词法分析工具包，具有中文分词和词性标注功能。

[Github地址](#)

StanfordCoreNLP

斯坦福NLP组的开源，支持python接口。

[Github地址](#)

HanLP

HanLP是一系列模型与算法组成的NLP工具包，由大快搜索主导并完全开源，目标是普及自然语言处理在生产环境中的应用。

[Github地址](#)

NLTK

NLTK是一个高效的Python构建的平台,用来处理人类自然语言数据。

[Github地址](#)

SpaCy

工业级的自然语言处理工具，遗憾的是不支持中文。

Gihub地址 | 官网

代码已上传至

本文转自 公众号 AI小白入门, [原文地址](#)



easyAI – 产品经理的 AI 知识库
<https://easyai.tech>



公众号: easyai-tech

命名实体识别 – NAMED-ENTITY RECOGNITION | NER



一文看懂命名实体识别

easyai

什么是命名实体识别?

命名实体识别（Named Entity Recognition，简称NER），又称作“专名识别”，是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等。简单的讲，就是识别自然文本中的实体指称的边界和类别。

[百度百科详情](#) | [维基百科详情](#)

命名实体识别的发展历史

NER一直是NLP领域中的研究热点，从早期基于词典和规则的方法，到传统机器学习的方法，到近年来基于深度学习的方法，NER研究进展的大概趋势大致如下图所示。



阶段 1：早期的方法，如：基于规则的方法、基于字典的方法

阶段 2：传统机器学习，如：HMM、MEMM、CRF

阶段 3：深度学习的方法，如：RNN – CRF、CNN – CRF

阶段 4：近期新出现的一些方法，如：注意力模型、迁移学习、半监督学习的方法

4类常见的实现方式

NER 的实现方式



监督学习



半监督学习



无监督学习



混合方法

早期的命名实体识别方法基本都是基于规则的。之后由于基于大规模的语料库的统计方法在自然语言处理各个方面取得不错的效果之后，一大批机器学习的方法也出现在命名实体类识别任务。宗成庆老师在统计自然语言处理一书粗略的将这些基于机器学习的命名实体识别方法划分为以下几类：

有监督的学习方法：这一类方法需要利用大规模的已标注语料对模型进行参数训练。目前常用的模型或方法包括隐马尔可夫模型、语言模型、最大熵模型、支持向量机、决策树和条件随机场等。值得一提的是，基于条件随机场的方法是命名实体识别中最成功的方法。

半监督的学习方法：这一类方法利用标注的小数据集（种子数据）自举学习。

无监督的学习方法：这一类方法利用词汇资源（如WordNet）等进行上下文聚类。

混合方法：几种模型相结合或利用统计方法和人工总结的知识库。

值得一提的是，由于深度学习在自然语言的广泛应用，基于深度学习的命名实体识别方法也展现出不错的效果，此类方法基本还是把命名实体识别当做序列标注任务来做，比较经典的方法是LSTM+CRF、BiLSTM+CRF。

NER 的相关数据集

数据集	简要说明	访问地址
电子病例测评	CCKS2017开放的中文的电子病例测评相关的数据	测评1 测评2

音乐领域	CCKS2018开放的音乐领域的实体识别任务	CCKS
位置、组织、人...	这是来自GMB语料库的摘录，用于训练分类器以预测命名实体，例如姓名，位置等。	kaggle
口语	NLPCC2018开放的任务型对话系统中的口语理解评测	NLPCC
人名、地名、机构、专有名词	一家公司提供的数据集,包含人名、地名、机构名、专有名词	boson

相关工具推荐

工具	简介	访问地址
Stanford NER	斯坦福大学开发的基于条件随机场的命名实体识别系统，该系统参数是基于CoNLL、MUC-6、MUC-7和ACE命名实体语料训练出来的。	官网 GitHub地址
MALLET	麻省大学开发的一个统计自然语言处理的开源包，其序列标注工具的应用中能够实现命名实体识别。	官网
Hanlp	HanLP是一系列模型与算法组成的NLP工具包，由大快搜索主导并完全开源，目标是普及自然语言处理在生产环境中的应用。支持命名实体识别。	官网 GitHub地址
NLTK	NLTK是一个高效的Python构建的平台,用来处理人类自然语言数据。	官网 GitHub地址
SpaCy	工业级的自然语言处理工具，遗憾的是不支持中文。	官网 GitHub地址
Crfsuite	可以载入自己的数据集去训练CRF实体识别模型。	文档 GitHub地址

本文转载自公众号 AI 小白入门，[原文地址](#)

欢迎关注我们的公众号：产品经理的 AI 知识库（微信号：easyai-tech）



easyAI – 产品经理的 AI 知识库
<https://easyai.tech>



公众号: easyai-tech