# 自然语言处理大实验

### 实验目的

* 了解自然语言处理领域的常见任务

自然语言处理有非常多的子领域，且很多任务都尚有巨大的提升空间。本实验的目的是了解一些最常见 的NLP 任务，阅读当前研究论文和相关数据集，并提供基准数据集和可能感兴趣任务的当前研究的论文推荐，作为未来研究的基础。

* 通过实践对自然语言处理形成相对完整而系统的认识

掌握自然语言处理领域的基础理论和方法，从数据获取和预处理、定义神经网络、定义损失函数和优化器、训练网络和测试网络，动手实践解决任务的整个流程。并从任务本身的特点出发思考如何搭建解决问题的模型方案。

### 任务一：信息抽取

* **任务介绍**

信息抽取（information extraction），简称IE，即给出一段文本，抽取出其中包含的实体和关系，并从数据库中查找的该实体所对应的其他关系。包含命名实体识别、实体链接、关系抽取等技术。命名实体识别的任务是找到文本中提到的每个命名实体，并标记其类型。一旦提取了文本中的所有命名实体，就可以将它们链接到与实际实体相对应的集合中。关系抽取则是发现和分类实体之间的语义关系。

* **数据集推荐**

1. NYT-10

NYT-10关系抽取数据集包含了53个关系，522043个句子实例， 其文本来源于纽约时报New York Times所标注的语料，命名实体是通过 Stanford NER 工具并结合 Freebase 知识库进行标注的。命名实体对之间的关系是链接和参考外部的Freebase知识库中的关系，结合远监督方法所得到的。

数据获取方式：<https://github.com/thunlp/OpenNRE/tree/master/benchmark>，下载并运行download\_nyt10.sh

1. Wiki80

对于给定了的句子和两个做了标注的名词，从给定的关系清单中选出最合适的关系。

数据集中一共包含80中关系，经统计各个关系个数均为700，合计56000个样本。数据获取方式：<https://github.com/thunlp/OpenNRE/tree/master/benchmark>，下载并运行download\_wiki80.sh

1. 人民日报标注语料库

人民日报标注语料库是对人民日报1998年上半年的纯文本语料进行了词语切分和词性标注制作而成的，严格按照人民日报的日期、版序、文章顺序编排的。文章中的每个词语都带有词性标记。

数据获取方式：

<https://github.com/AlanConstantine/PeoplesDailyNER/blob/master/199801_people_s_daily.txt>

1. SemEval2010

SemEval数据集来自于2010年的国际语义评测大会中Task 8，对于给定句子和两个做了标注的名词，从给定的关系清单中选出最合适的关系。

数据获取方式：<https://github.com/thunlp/OpenNRE/tree/master/benchmark>，下载并运行download\_semeval.sh

* **数据集实例**

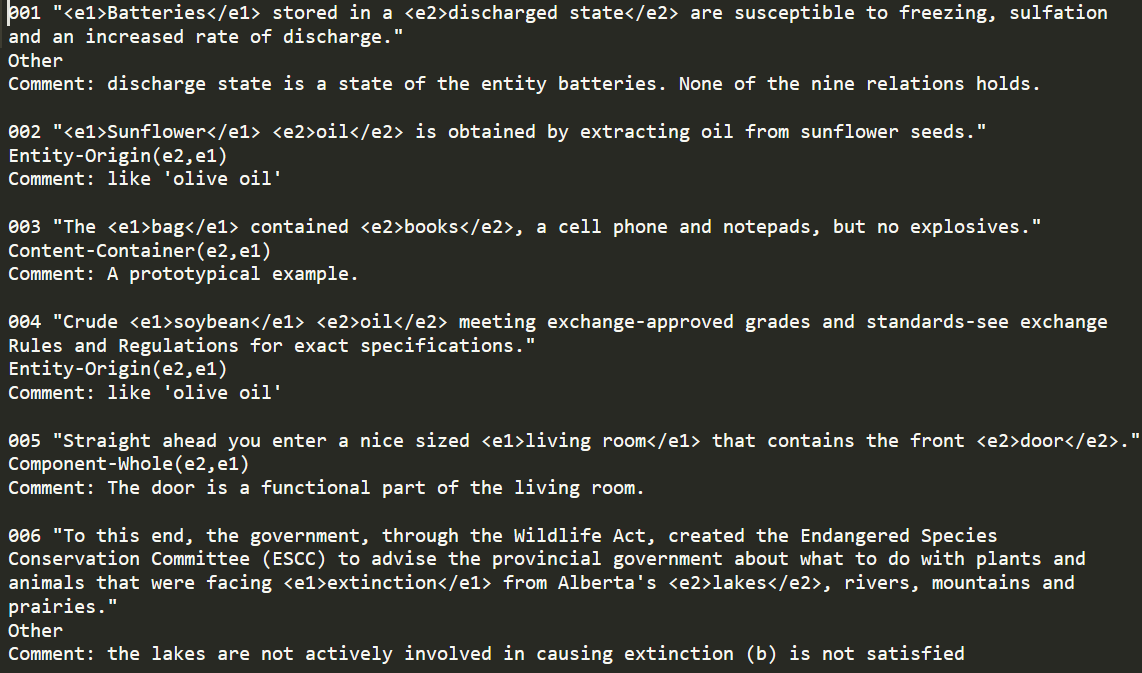


图 1 SemEval2010 数据集实例

* **实验要求**

1. 实现实体识别与关系抽取算法
2. 将抽取到的数据存入图数据库中（neo4j）
3. 利用GUI或者网页进行demo演示
4. 输入一条文本，抽取出其中的实体与关系，指定一个实体，可以从数据库中获取该实体涉及的所有关系

* **文献推荐**

1. [Chinese NER Using Lattice LSTM](https://aclanthology.org/P18-1144/)
2. [Relation Classification via Convolutional Deep Neural Network](https://aclanthology.org/C14-1220/)
3. [Attention-Based Convolutional Neural Network for Semantic Relation Extraction.](https://aclanthology.org/C16-1238/)
4. [An Attention-Based Model Using Character Composition of Entities in Chinese Relation Extraction](https://www.mdpi.com/2078-2489/11/2/79)
5. [Enriching Pre-trained Language Model with Entity Information for Relation Classification](https://arxiv.org/abs/1905.08284)
6. [Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions](https://www.semanticscholar.org/paper/Distant-Supervision-for-Relation-Extraction-with-Ji-Liu/b8da823ad81e3b8e5b80d82f86129fdb1d9132e7)

### 任务二：摘要生成

* **任务介绍**

给定一篇文档，生成该文档的摘要。目前主流的文本摘要自动生成有两种方式，一种是抽取式（extractive），另一种是生成式（abstractive）。抽取式，顾名思义就是按照一定的权重，从原文中寻找跟中心思想最接近的一条或几条句子作为文章的摘要，基本上是对原文重要部分的汇总，摘取原文措辞不变。而生成式则是在原文被完全理解的基础上，用全新的词概括，代表原文的主旨，该任务要求采用生成模型完成。

* **数据集推荐**
  1. LCSTS

LCSTS是一个大规模、高质量中文短文本摘要数据集。数据集中包含了200万真实的中文短文本数据和每个文本作者给出的摘要。同时也手动标注了10666份文本的摘要。数据获取方式：<https://github.com/yamonc/New-Pytorch-Chinese>(0.数据预处理)

1. CNN/Daily Mail

CNN/Daily Mail作为单文本摘要语料库，每篇摘要包含多个摘要句。数据集最初是从美国有限新闻网（CNN）和每日邮报网（Daily Mail）收集的约100万条新闻数据作为机器阅读理解语料库。后来进行简单改动，形成用于单文本生成式摘要的语料库。将每篇新闻的要点按原文中出现的顺序组成多句的摘要，每个要点看成是一个句子。数据获取方式：<https://github.com/abisee/cnn-dailymail>

* **数据集实例**

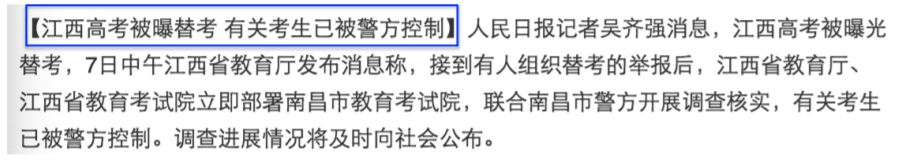


图 2 LCSTS数据实例

* **实验要求**

1. 实现摘要生成算法
2. 利用GUI或者网页进行demo演示
3. 输入一段文档，生成对应的摘要（标题）

* **文献推荐**

1. [A Hierarchical Structured Self-Attentive Model for Extractive Document Summarization (HSSAS).](https://arxiv.org/abs/1805.07799)
2. [Hierarchical Learning for Generation with Long Source Sequences](file:///D:\\Tencent%20Files\\3526127072\\FileRecv\\2.%09Hierarchical%20Learning%20for%20Generation%20with%20Long%20Source%20Sequences)
3. [PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization](https://arxiv.org/abs/1912.08777)
4. [A Neural Attention Model for Abstractive Sentence Summarization](https://arxiv.org/abs/1509.00685)
5. [BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension](https://arxiv.org/abs/1910.13461)

### 任务三：对话机器人

* **任务介绍**

对话机器人是经由对话或文字进行交谈的程序。对于一段输入文本，对话机器人能够生成相应的回复。构建对话机器人的模型通常可以分为基于检索的模型和生成模型。基于检索的模型，使用预定义回复的数据库和某种启发式推理来根据输入和上下文选择适当的响应，不会产生任何新的文本，只是从固定的集合中选择一个回复。而生成式模型不依赖于预定义的回复，完全从零开始生成新的回复。

* **数据集推荐**

1. ConvAI2

带个性的persona-chat对话数据，对话目标是了解对方。数据集来自于众包人员之间的真实对话。这些众包人员随机配对并被要求扮演给定的persona信息（persona信息随机分配，由另一组众包人员创建）。数据获取方式：

<http://parl.ai/downloads/convai2/convai2_fix_723.tgz>

1. Pchatbot

一个大规模对话数据集，包含分别从微博和司法论坛收集的公开 post-response 数据。并且为帖子和回复提供匿名用户 ID 和时间戳。数据获取方式：

<https://github.com/qhjqhj00/SIGIR2021-Pchatbot>

1. Empathetic Dialogue

Empathetic Dialogues数据集包含约33090个对话，每个对话仅包含少量句子，并被分类为不同的情感状况。数据获取方式：

<http://parl.ai/downloads/empatheticdialogues/empatheticdialogues.tar.gz>

1. Wizard of Wikipedia

该对话数据集属于开放域对话系统，一个对话者随机选择一个初始话题，对话双方可以在此基础上进行对话，但在对话过程中话题也可以拓展。对话双方的角色是不同的，分为wizard和apprentice。数据获取方式：

<http://parl.ai/downloads/wizard_of_wikipedia/wizard_of_wikipedia.tgz>

* **数据集实例**

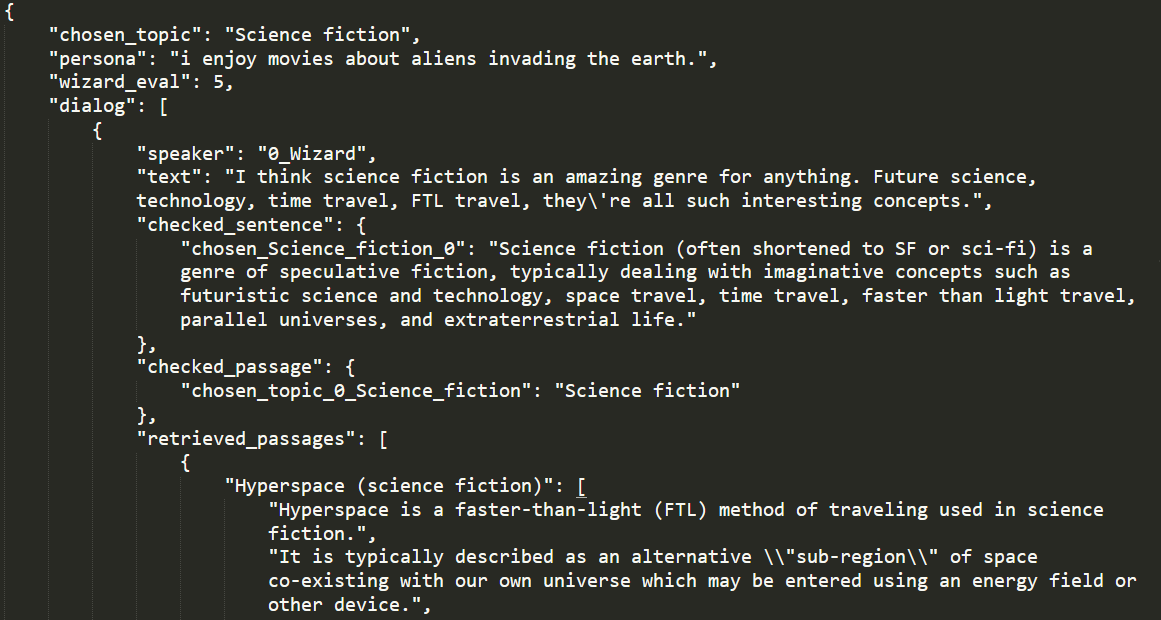


图 3 Wizard of Wikipedia数据实例

* **实验要求**

1. 实现对话生成算法
2. 利用GUI或者网页进行demo演示
3. 输入一句话，给出相应的回复

* **文献推荐**

1. [Personalizing Dialogue Agents: I have a dog, do you have pets too](https://arxiv.org/abs/1801.07243)
2. [Recipes for building an open-domain chatbot](https://arxiv.org/abs/2004.13637)
3. [You Impress Me: Dialogue Generation via Mutual Persona Perception](https://arxiv.org/abs/2004.05388)
4. [PLATO-2: Towards Building an Open-Domain Chatbot via Curriculum Learning](https://arxiv.org/abs/2006.16779)
5. [Multiresolution Recurrent Neural Networks: An Application to Dialogue Response Generation](https://arxiv.org/abs/1606.00776)

### 任务四：名词短语数据分析

* **任务介绍**

给定一篇文档，挖掘文章中包含的名词短语以及这些名词之间的语义关系。此处涉及到的语义关系包含三大类：1)共指关系：别称、指代、近义词；2)系统关系：类属、要素、领有、邻近；3)蕴含关系：逻辑、属性、逆向。每个小类的解释和示例如表1所示。可以使用基于规则的方法、基于无监督学习的方法和基于有监督学习的方法来实现名词短语数据分析。

表 1 语义关系解释及举例



* **数据集**

Chinese Treebank 9.0

该数据集由中文新闻专线、政府文件、杂志文章、各种广播新闻和广播对话节目、网络新闻组、网络日志、论坛、聊天信息和转录的对话电话讲话中的注释和解析文本组成。9.0版本增加了更多的带注释的网络数据和两种新类型——聊天信息和转录的对话电话语音。数据获取方式：从群文件处可下载

* **数据集实例**

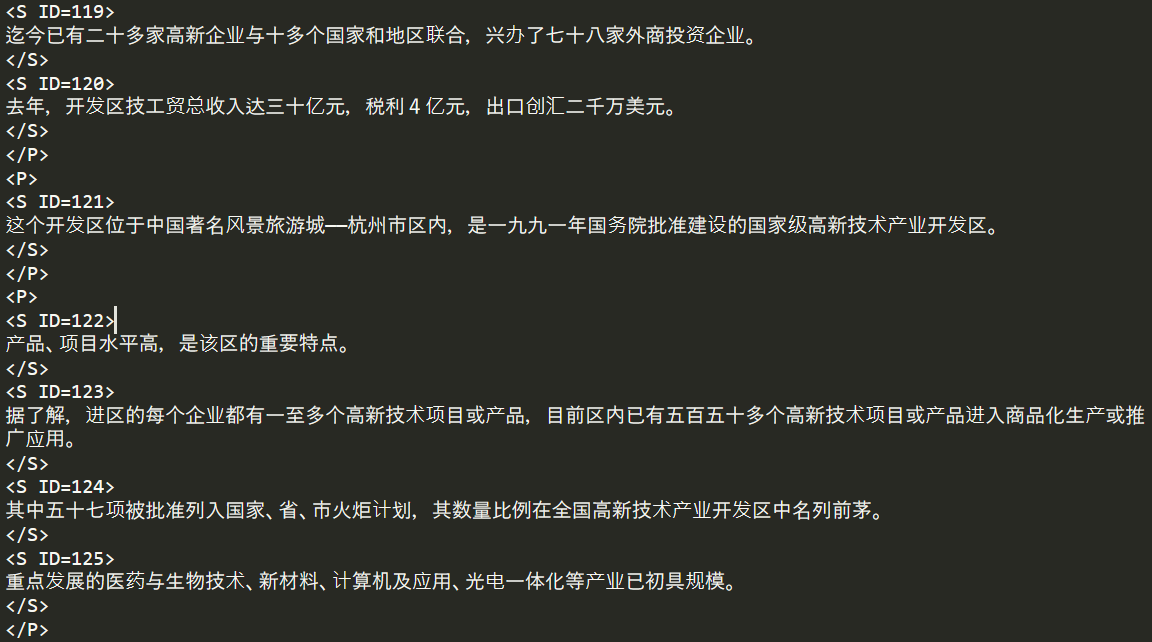


图 4 Chinese Treebank9.0数据实例

* **实验要求**

1. 尽可能找出语料中包含的名词
2. 尽可能每两个名词间对应的关系
3. 统计并分析名词及其关系

### 提交时间

**1月20号截止**

一个小组由一名同学负责，提交实验报告以及相关代码（不包括模型和数据）到3648311320@qq.com，邮件以及实验报告命名方式为第X组——实验名称，实验报告封面请注明实验名称、组号、小组成员。

### 实验要求

* 完成选定实验任务
* 实验报告中请注明小组分工
* 良好的代码风格
* 完整的实验报告
* 完整的项目演示，清晰的实验结果截图
* 可以组队完成，队伍不超过4人