学术文本中的关键词提取与实体链接

Project Report for EE447

陈奕廷 Yiting Chen
517021910169
sjtucyt@sjtu.edu.cn

2020/6

1 任务简介

随着论文总量逐年增加，在海量的文献中快速检索与理解文本内容，对文本内容进行信息凝练成为一项重要目标。因此可以利用自然语言处理的技术，实现如关键词提取，实体链接等辅助阅读文本的手段。接下来我先简要介绍关键词提取与实体链接的概念。

1.1 关键词提取

关键词提取，顾名思义即通过统计或机器学习等手段将一段文本中的关键词提取出来。因为关键词提取也相当于对文本中的词按照其重要性排序，因此可以将关键词提取的问题转化为词排序问题。

1.2 实体链接

文本中出现的名词作为实体，在语义上可能与多个同名实体相混淆（如华盛顿可指历史人物也可指地名）。实体链接所要实现的即将文本中出现的

实体与已有实体库中的特定实体相链接，起到消岐的作用。

2 相关工作

2.1 自然语言处理

自然语言处理，作为人工智能的一个重要研究方向，是人工智能，计算机科学，语言学所关注的计算机与自然语言之间相互作用的领域，近年来受到比较多的关注。许多研究依托循环神经网络如长短期记忆网络（LSTM）以及诸多统计与数据挖掘技术在命名实体识别，词性标注等自然语言理解方向取得了很大的进步。

2.2 语料库语言学

语料库语言学主要研究机器可读的自然语言文本的句法标注、句法语义分析、语言定量分析、作品风格分析以及自然语言理解和机器翻译等领域的应用。语料库将原始的在语言的实际使用过程中真实出现过的文本进行收集、分析、整理与标注，作为语言学的研究材料。近年来，研究语料库语言学着重在多个层次如语音、构词、句法、语义及语用等层次的标注。

3 实现方法

3.1 关键词提取

关键词提取问题可以转化为对词的重要性排序问题，接下来我介绍几种从不同角度解决关键词提取的方法。

3.1.1 条件随机场（CRF）

作为机器学习领域的一个算法模型，条件随机场在给定随机变量 X 的条件下，建立关于随机变量 Y 的马尔可夫随机场。对于如文本处理之类的

线性链，可以建立一个线性链 CRF。首先定义特征函数 $f(i, b_i, a_i)$，它以单词与单词前后单词序列为输入。其中 i 表示文本中的第 i 个单词，b_i 与 a_i 分别表示单词前后的单词序列。则若共建模了 m 个特征函数，给每一个特征函数 f_j 加权 λ_j 后，可以对第 i 个单词评分：

$$score(i|s) = \sum_{j=1}^{m} \lambda_j f_j(i, a_i, b_i)$$

最后由特征函数至该单词为关键词的概率则：

$$Prob(i|s) = \frac{exp[score(i|s)]}{\sum_i' exp[score(i'|s)]}$$

从某种角度来说，这一方法类似于对序列数据所作的逻辑斯蒂回归。最终将得到的概率作为单词重要性的指标，即可将关键词提取出来。

在具体实现过程中，我采用的数据集是 Sem Eval 2017 Task 10。

3.1.2 TextRank 算法

启发于之前 Page Rank 的算法，TextRank 通过图算法来解决文本中单词排序的问题。实现的过程分为两步。

第一步，构建一个以文本内单词为节点的图，若两个单词在长度为 m 个单词的窗口内共现（即两个单词间的单词小于 $m-1$ 个）则这两个单词的节点间存在无向边。可以理解为将单词间的联系以图中边的形式建模。

第二步，对于已经构建好的图，要计算词的重要性主要依据两条原则：

- 一个词与越多的词有联系（有无向边相连）则该词越重要。
- 一个词与越重要的词有联系，则该词越重要。

具体实现步骤为先为每个节点（词）分配同样的权重，然后根据

$$PR(i) = \sum_{ij \in E} \frac{PR(j)}{N_j}$$

Genre group	Category	Content of category	No. of texts
Press (88)	A	Reportage	44
	B	Editorial	27
	C	Review	17
General Prose (206)	D	Religion	17
	E	Skills, trades and hobbies	36
	F	Popular lore	48
	G	Belles lettres, biographies, essays	75
	H	Miscellaneous	30
Learned (80)	J	Science	80
Fiction (126)	K	General fiction	29
	L	Mystery and detective Fiction	24
	M	Science fiction	6
	N	Adventure and Western	29
	P	Romance and love story	29
	R	Humor	9
Total			**500**

图 1: Category in FLOB

对权重进行迭代，直至收敛。其中 $PR(i)$ 表示第 i 个单词的权重，$ij \in E$ 表示 i 与 j 间有有边相连，N_j 表示与 j 单词相连单词的个数。

因为算法复杂度在 $O(NM^2)$ 同时囿于电脑性能，实际实现时对论文分段提取关键词。

3.1.3 基于语言学语料库的关键词提取

语料库语言学中，为了研究多种不同类型的文本的句法特点，用词偏好，往往使用语料库中其他文本作为背景文本，通过词频等统计学指标研究目标领域文本的特点。同理，也可以将词频等统计学手段用于关键词提取。我挑选了分别为美国英语语料库与英国英语语料库的 FROWN 和 FLOB 中的学术文本（J 类）作为关键词提取的背景文本，将目标文本的词频与背景文本的词频进行比对，可以认为，在目标文本中词频显著高于背景文本

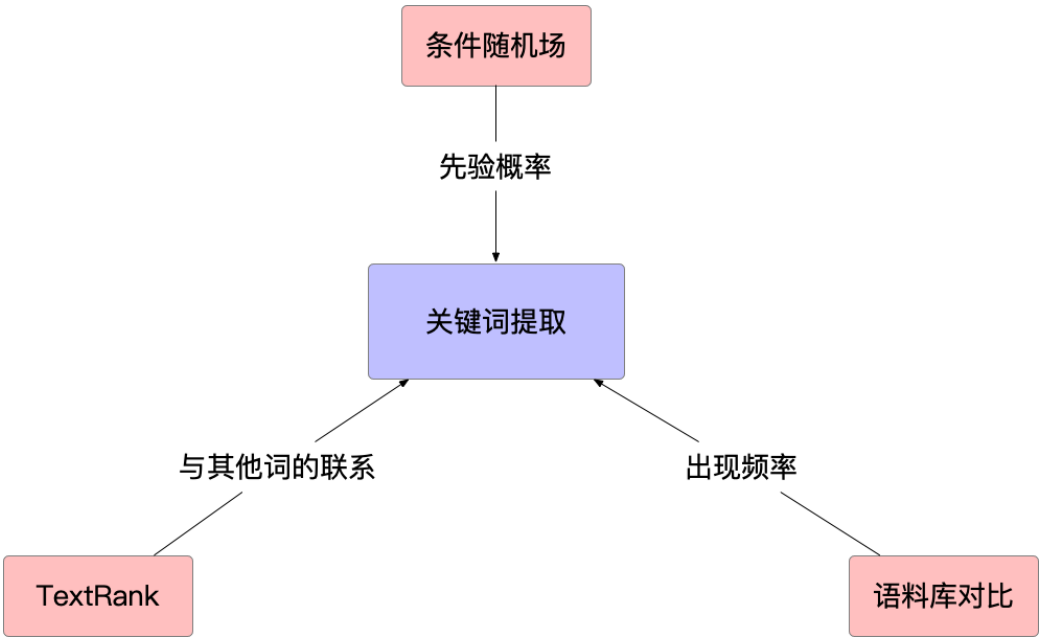


图 2: 三个角度实现关键词提取

的词即为关键词。

3.1.4 多级关键词提取

从不同角度用不同方式度量一个词时往往会呈现不同重要性。条件随机场从先验概率的角度，根据已有数据集中的标记对文本中的词进行重要性判定，TextRank 从词与其他同一文本中词的联系的角度对词的重要性进行排序。而统计学，语料库语言学的衍生方法则从词的出现频率这个角度对词的重要性进行判定。

不同角度均有其物理意义，结合以上三种从不同角度不同手段的关键词提取方式，我尝试给予提取出的关键词一个综合关键性评分并进行分级。将三种手段得出的词重要性分数归一化后求和，可得出一个最终关键性评分。

同时按一个词同时出现在几个手段结果中的前 10% 对关键词分级，如：若一个词在三种方法测量均为前 10% 则为 1 级关键词，仅出现在两个方法

测量的前 10% 则为 2 级关键词，仅出现在一个方法的前 10% 为 3 级关键词。

3.2 实体链接

实体链接为将文本中的实体与实体库中实体相链接。主要可以分成三步，第一步分割文本，找出文本中的实体。第二步对各种实体消岐。第三步，将所有可能与文本中实体链接的实体按可能性排序。其中最主要的任务即为对文本中出现的实体进行语义上的消岐。

目前主流使用的实体库有维基百科的词条整理而来。其中选择维基百科有两个理由：其一，维基百科中词条数量众多，基本覆盖所有可能出现的实体。其二，词条与词条间有链接，这使得通过实体间的联系实现消岐成为可能。

为方便描述，定义可以用以指向一个或多个实体的词为锚点。对于一篇文本中的锚点，消岐的最主要想法是，一个锚点可能指向的实体与同一文本内其他锚点所指向实体的联系越多，则越有可能是与该锚点链接的实体。

对于文本内锚点 a 通过其他锚点对其投票的方式来确定所链接的实体，其中锚点 b 对其中一个实体 p_a 所投的票可以写作：

$$Vote_b(p_a) = \frac{\sum_{p_b \in Pg(b)} rel(p_a, p_b) Pr(p_b|b)}{|Pg(b)|}$$

其中 p_a 和 p_b 分别表示可能与 a 或 b 链接的实体，$Pg(b)$ 表示锚点 b 可能指向的所有实体集合。$rel(p_a, p_b)$ 表示实体 p_a 和 p_b 间的联系，若有链接相连为 1，没有为 0。$Pr(p_b|b)$ 则表示锚点 b 指代实体 p_b 的概率。

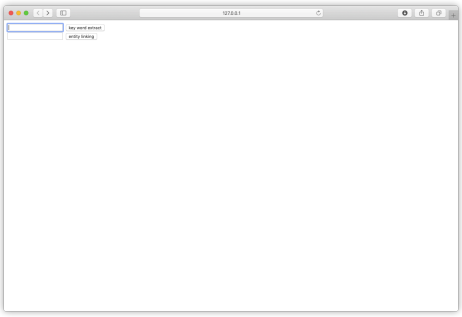
将不同实体获得的投票作为排序的依据，可以做到消岐，进行实体链接。

4 结果展示

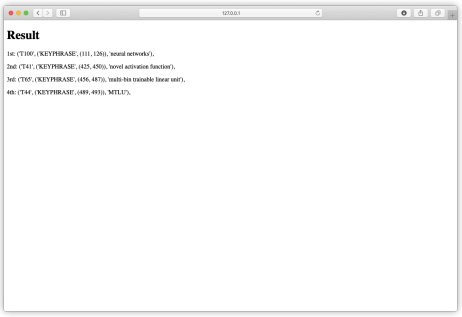
4.1 网页实现

为了分别展现关键词提取与实体链接的实现效果，我编写了一个网页来展示所实现的文本关键词提取与实体。

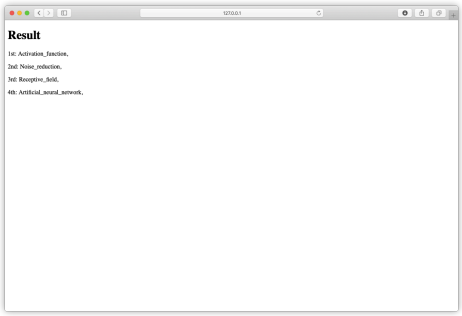
在搭建网络方面，我本人之前并没有搭建网络的经验。因此选择了 python 编程的 Django 框架。这个网页极其简单，几乎不分前端与后端。主要实现的功能是对文本框内的文本分别进行关键词提取或实体链接，并展示最终提取出的关键词或文本中出现的实体库中的实体。



(a) Home page



(b) keyword-extraction



(c) entity-linking

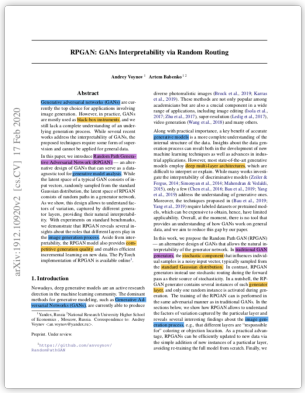
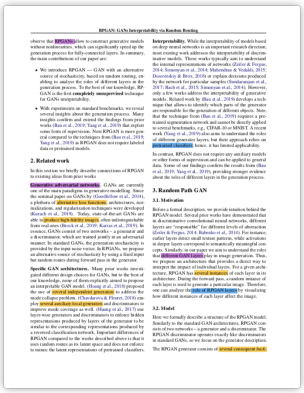
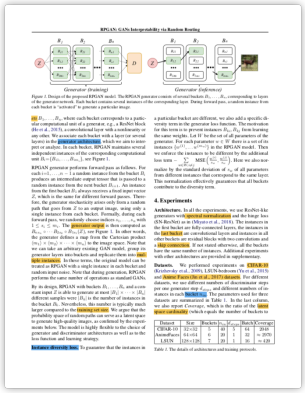
图 3: 网页效果截图

如图 3所示，这是我使用网站对一篇有关去噪论文的摘要进行关键词

提取与实体链接的截图。

4.2 论文关键词提取

对整篇论文的关键词提取，囿于电脑性能，我实行分段提取关键词。

(a) Page 1 (b) Page 2 (c) Page 3

图 4: 论文关键词提取效果

图 4 是一篇关于对抗生成网络解释性的论文。通过关键词提取后，紫色、蓝色、黄色分别表示 1 级关键词，2 级关键词，3 级关键词。我展示了这篇论文的前三页，后面实验部分论文内图表过多，暂时不展示。可以看到通过三种方法结合提取出的 1 级与 2 级关键词较好的体现了论文主旨，可以有效起到辅助阅读的功能

4.3 其他结果展示

其他结果包括词云图等

如图 5，采用网页提取的关键词所制成的词云图。越大的单词代表越关键。

其余结果包括网页实操已经在课程汇报时展示过。

图 5: 词云图

5 总结

本次课程 project 中，成功实现了学术文本的关键词提取与实体链接。其中在关键词提取部分，从三个不同的角度 (先验概率，词与词联系，词频) 分别实现了关键词提取，并将其综合成一个指标对文本内词的重要性进行了衡量与排序并对关键词分级。起到了一定的辅助阅读功能。

在结果展示方面，开发实现了一个功能性网页，并对真实论文实现了关键词标注。并通过词云图等制作，尽量是结果展示多样化。

最后感谢傅老师和助教一学期的辛勤教导和帮助

感谢来给我们讲过课的学长学姐