EE447

学者主页
关键信息提取

李辰轩　　林珈漫　　丁晨诗

01.结果展示+网页搭建

02.数据获取+文本处理

03.关键词抽取

04.实体抽取识别

结果展示

网页搭建

FastAPI + bootstrap

FastAPI + Bootstrap

Web后端框架：FastAPI

轻量级异步并发Web框架，基于Python3.6+

特点：

- **Fast**: Very high performance, on par with **NodeJS** and **Go** (thanks to Starlette and Pydantic). One of the fastest Python frameworks available.
- **Fast to code**: Increase the speed to develop features by about 200% to 300%. *
- **Fewer bugs**: Reduce about 40% of human (developer) induced errors. *
- **Intuitive**: Great editor support. Completion everywhere. Less time debugging.
- **Easy**: Designed to be easy to use and learn. Less time reading docs.
- **Short**: Minimize code duplication. Multiple features from each parameter declaration. Fewer bugs.
- **Robust**: Get production-ready code. With automatic interactive documentation.
- **Standards-based**: Based on (and fully compatible with) the open standards for APIs: OpenAPI [↪] (previously known as Swagger) and JSON Schema [↪].

Web前端框架： Bootstrap

基本结构：Bootstrap 提供了一个带有网格系统、链接样式、背景的基本结构。
CSS：Bootstrap 自带以下特性：全局的 CSS设置、定义基本的 HTML 元素样式、可扩展的 class，以及一个先进的网格系统。
组件：Bootstrap 包含了丰富的可重用的组件。
JavaScript 插件：Bootstrap包含了十几个自定义的 jQuery 插件。
利用这些组件、插件，可以帮助我们更快速、容易地搭建一个功能完备的网站

数据获取+文本处理

数据获取

- 学者主页地址获取：课程参考数据库资源大多为国内外学者论文、会议等信息，无学者主页地址等信息， 且考虑爬取Google国外学者的学者主页有困难，所以目前建立交大计算机系学者主页数据库，分析其学者主页提取关键信息。
- http://www.cs.sjtu.edu.cn/Faculty.aspx 为计算机系教师名录官网，爬取教师个人信息网站可获取其个人主页地址和一些其他基本信息，建立数据库。

page_url	peopleDetail_url	scholar_field	scholar_id	scholar_name
http://basics.sjtu.edu.cn/~yuxi/	http://www.cs.sjtu.edu.cn/PeopleDetail.aspx?id=103	高可靠软件与理论研究所	1	傅育熙
http://www.cs.sjtu.edu.cn/~kzhu/	http://www.cs.sjtu.edu.cn/PeopleDetail.aspx?id=63	高可靠软件与理论研究所	2	朱其立
http://basics.sjtu.edu.cn/~xiaoju/	http://www.cs.sjtu.edu.cn/PeopleDetail.aspx?id=107	高可靠软件与理论研究所	3	董笑菊
NULL	http://www.cs.sjtu.edu.cn/PeopleDetail.aspx?id=254	高可靠软件与理论研究所	4	龙环
http://basics.sjtu.edu.cn/~dominik/	http://www.cs.sjtu.edu.cn/PeopleDetail.aspx?id=304	高可靠软件与理论研究所	5	Dominik Scheder
http://jhc.sjtu.edu.cn/~hongfeifu/	http://www.cs.sjtu.edu.cn/PeopleDetail.aspx?id=388	高可靠软件与理论研究所	6	符鸿飞
http://chihaozhang.com/	http://www.cs.sjtu.edu.cn/PeopleDetail.aspx?id=406	高可靠软件与理论研究所	7	张驰豪
NULL	http://www.cs.sjtu.edu.cn/PeopleDetail.aspx?id=111	并行与分布计算研究所	8	陈贵海
http://www.cs.sjtu.edu.cn/~hbguan/	http://www.cs.sjtu.edu.cn/PeopleDetail.aspx?id=102	并行与分布计算研究所	9	管海兵
http://www.cs.sjtu.edu.cn/~guo-my/	http://www.cs.sjtu.edu.cn/PeopleDetail.aspx?id=100	并行与分布计算研究所	10	过敏意

学者网页

Yanyan SHEN (沈艳艳)

Tenure-track Associate Professor

SEIEE Building #03-528
Data Driven Software Technology (DDST) Laboratory
Department of Computer Science and Engineering
Shanghai Jiao Tong University
Email: shenyy AT sjtu DOT edu DOT cn

Welcome to visit our DDST lab website!

About Me

I am a tenure-track associa
Jiao Tong University (SJTU
2010, and got my doctoral
at AT&T Shannon Lab (duri

Weinan ZHANG

张伟楠

Associate Professor

John Hopcroft Center for Computer Science
Department of Computer Science & Engineering
Shanghai Jiao Tong University

307 Yifu Building, 800 Dongchuan Road
Shanghai, 200240, China

Email: wnzhang [AT] sjtu.edu.cn

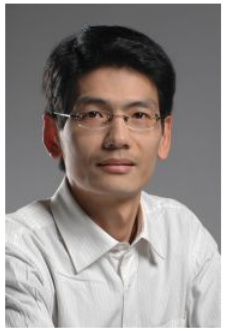
Weinan Zhang is now a tenure-track associate professor at Shanghai Jiao Tong University. His research interests include (multi-agent) reinforcement learning, deep learning and data science with various real-world applications of recommender systems, search engines, text mining & generation, knowledge graphs, game AI etc. He has published over 80 research papers on international conferences and journals and has been serving as a (senior) PC member at ICML, NeurIPS, ICLR, KDD, AAAI, IJCAI, SIGIR etc. and a reviewer at JMLR, TOIS, TKDE, TIST etc. He was granted as Top-20 Rising Stars in 2016 by Microsoft Research, Best Paper Honorable Mention Award in SIGIR 2017, ACM Shanghai Rising Star Award 2017, Alibaba DAMO Young Scholar Award 2018 and the Best Paper Award in DLP-KDD Workshop 2019.

Weinan earned his Ph.D. from University College London in 2016 and B.Eng. from ACM Class of Shanghai Jiao Tong University in 2011. He was an intern at MediaGamma, Microsoft Research, Google and DERI.

Prospective Ph.D. students: I am looking for outstanding and highly motivated Ph.D. students to work together on multi-agent reinforcement learning, model-based reinforcement learning, graph deep learning and various data mining topics. Please email me with your CV and transcripts.

Notices:

- Partipate our challenge on mobility intervention for epidemics held in conjunction with our KDD'20 workshop!



Home
Publications
Patents
Teaching

Associate Professor, Ph.D.

Department of Computer Science and Engineering,
Shanghai Jiao Tong University

Office:
3-535 SEIEE Building, / Floor 3, building 4, software building,
800 Dongchuan RD. Minhang District,
Shanghai 200240, P.R. China

Laboratory:
Emerging Parallel Computing Center.
High-performance Data Processing Laboratory.

Email: yshen@cs.sjtu.edu.cn

Research Interests

- Cloud computing and distributed systems;
- Deep learning for image processing;
- Computer networking, with an emphasis on SDN.

Recent Work

- Deep learning on object detection and identification (e.g. face detection /identification /liveness detection /other feature extractions, pose estimation, bird detection and

News

2 May 2020
I co-organize 2nd Workshop on Deep Learning Practice for High-Dimensional Sparse Data with KDD 2020.

1 May 2020
I co-organize SIGIR 2020 Workshop on Deep Reinforcement Learning for Information Retrieval.

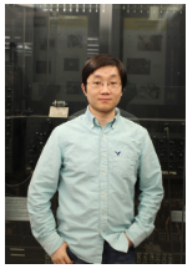
8 Aug 2019
Our paper wins the best paper award in DLP-KDD 2019.

3 Nov 2018
I will serve as a sponsorship chair in CIKM 2019.

1 Nov 2018
Two full papers and one demo paper accepted in AAAI 2019.

26 Sep 2018
I am honored to be granted Qingcheng Award from Alibaba.

12 May 2018
Two papers on RL and AutoML accepted in ICML 2018.



Dr. Quan Chen (陈 全)

Tenure-Track Associate Professor (特别研究员)
Department of Computer Science and Engineering
Shanghai Jiao Tong University

Tel: +86-21-34207239
Office: Room 3-522, SEIEE Building
Address: No. 800, Dongchuan Road, Shanghai, China, 200240
Email: chen-quan [AT] cs [DOT] sjtu [DOT] edu [DOT] cn

Index

Research Interest

High performance computing, Parallel processing, Scheduling in various architectures;
Distributed System, Runtime System, Operating System;
Resource management in Datacenter.

数据处理

- 爬取学者主页，起初我们考虑学者主页格式内容非常多元化，没有统一结构，所以去掉了所有标签，只保留其文本数据内容处理。
- 得到文本内容后，我们过滤了中文页面、爬取失败的页面，去掉了文本内容中无意义和错误的换行符和空格以免影响分词和其他处理。之后尝试采用Rake算法对全文文本提取关键词，对文本进行实体识别等处理。但是发现针对全文处理的效果并不好，精确度不高，处理时间也更长。
- 重新考虑html网页源码的标签信息，利用<h...>标签等提取其标题信息，获得学者页面的结构特征，用以将文本数据分段。
- 查看分析了大量学者网页，发现了一些内容共性，将标题词义相似的段落内容归为一类，将文本段落内容按各分类分别存储。
- 对各类段落分别处理，降低处理难度，更高效也更有针对性，容易提取出我们想要的信息。

数据处理

关键信息：

- 我们提取了以下关键信息：
  - 学者联系方式：
    - 相关内容：电话、邮箱
    - 提取方法：正则项匹配
  - 研究方向：
    - 相关内容：学者主页研究方向文本内容，根据论文标题等生成的研究方向词云
    - 提取方式：正则项匹配， 关键词提取
  - 学者论文：
    - 相关内容：发表年份、标题、合作者及其相关信息
    - 提取方法：实体识别 + 可视化

关键词提取

关键词提取

处理海量的文本文件最关键的是要把用户最关心的问题提取出来。我们往往可以通过几个关键词窥探整个文本的主题思想。因此，在这里我们试图通过对学者主页的文本进行关键词的提取，来反映这个学者的科研方向等信息。

关键词提取的三种方法

有监督的关键词抽取算法

半监督的关键词提取算法

无监督的关键词抽取算法

无监督关键词提取算法

无监督关键词抽取算法可以分为三大类。

1 基于统计特征

2 基于词图模型

3 基于主题模型

基于统计特征的关键词提取

基于统计特征的关键词抽取方法的关键是采用什么样的特征值量化指标的方式，目前常用的有三类：

1 基于词权重的特征量化

基于词权重的特征量化主要包括词性、词频、逆向文档频率、相对词频、词长等。

2 基于词的文档位置的特征量化

这种特征量化方式是根据文章不同位置的句子对文档的重要性不同的假设来进行的。

3 基于词的关联信息的特征量化

词的关联信息是指词与词、词与文档的关联程度信息。

常用的特征值量化指标

关键词一般为名词

词性

位置信息

TF-IDF值

词频

词跨度

互信息

$$span_i = \frac{last_i - first_i + 1}{sum}$$

$$\mathrm{I(X;Y)} = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

TF-IDF算法

TFIDF的主要思想：

如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来作为关键词。TFIDF实际上是：TF * IDF，TF词频(Term Frequency)，IDF逆向文件频率(Inverse Document Frequency)。

TF表示词条在文档d中出现的频率。

某一特定词语的IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取以10为底的对数得到。

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad \mathrm{idf}_i = \lg \frac{|D|}{|\{j : t_i \in d_j\}|} \qquad \mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i$$

关键词提取流程

1 文本预处理

去除标签，分词，去除停用词，建立语料库

2 统计词频和逆文档频率

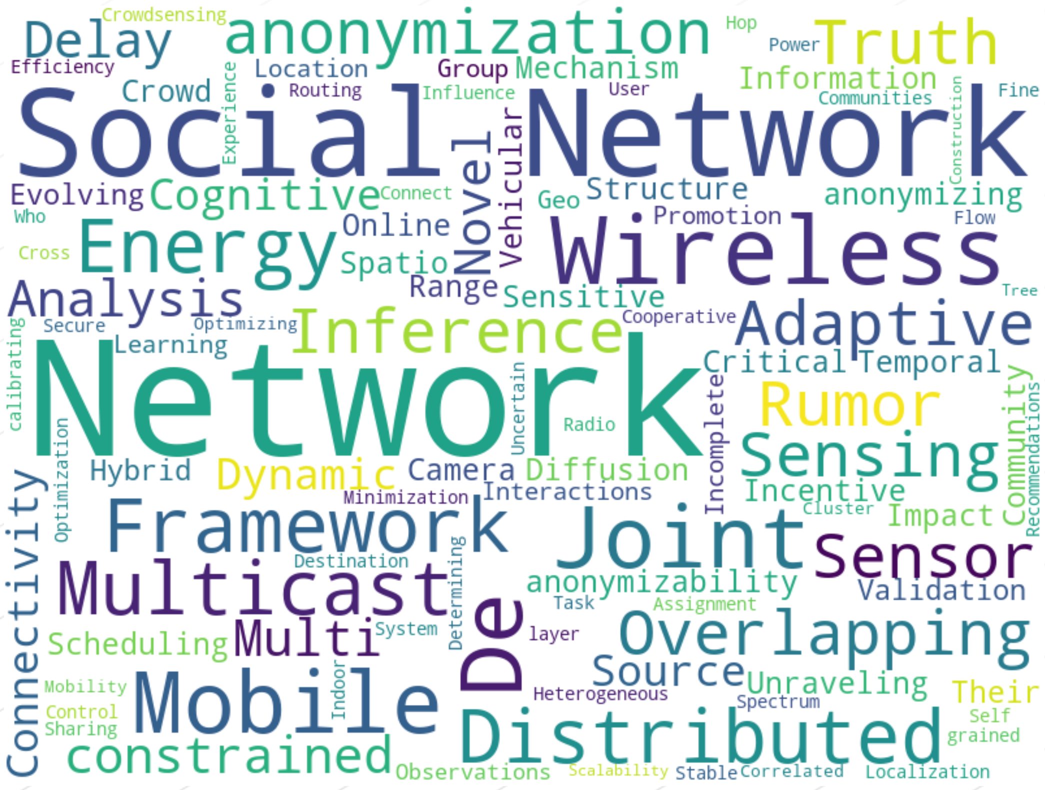
3 计算权重

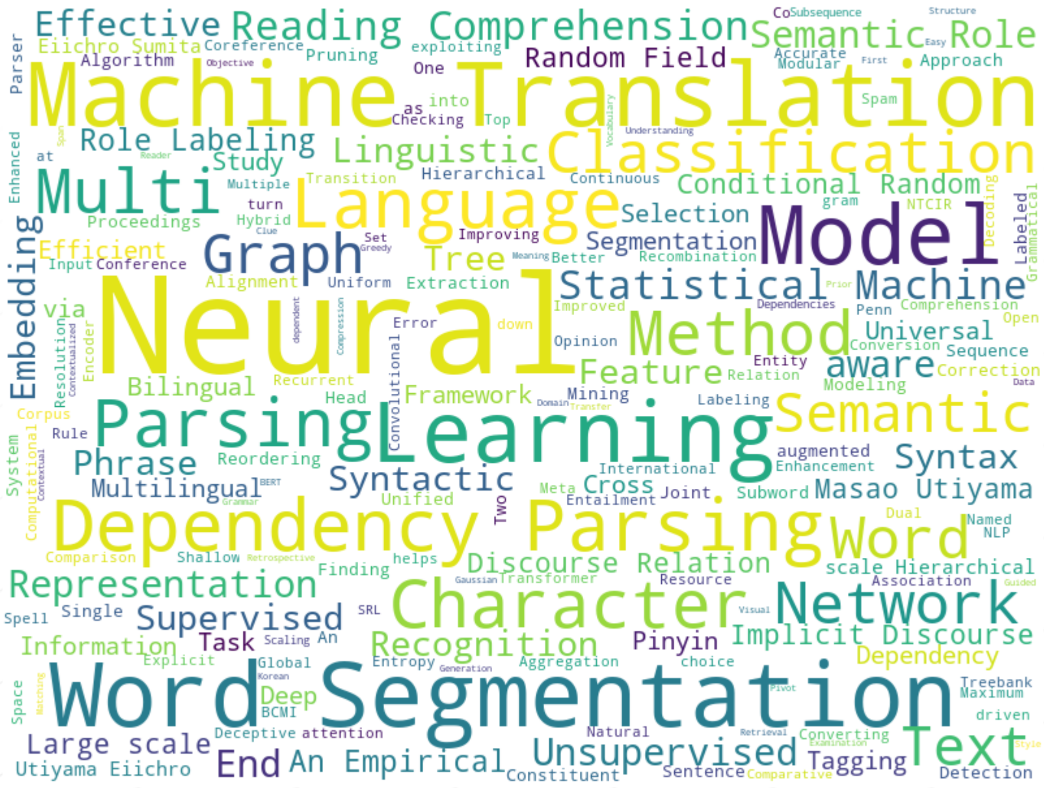
计算TF-IDF

4 可视化

归一化，生成词云

关键词提取结果


Social
Network
Wireless
Energy
Network
Analysis
Inference
Adaptive
Rumor
Sensing
Framework
Joint
Sensor
Multicast
Mobile
Distributed
Overlapping
Connectivity
constrained
Multi
De
Delay
anonymization
Truth
Crowd
Location
Group
Mechanism
Information
Cognitive
Online
Spatio
Novel
Vehicular
Structure
anonymizing
Range
Sensitive
Critical
Temporal
Dynamic
Camera
Diffusion
Incentive
Impact
Hybrid
Interactions
Minimization
anonymizability
Validation
Scheduling
Source
Unraveling
Their
Community
Recommendations
Incomplete
Uncertain
Evolving
Learning
Efficiency
Crowdsensing
Routing
Influence
User
Power
Hop
Communities
Fine
Geo
Promotion
Flow
Who
Cross
Secure
Optimizing
Cooperative
Tree
Radio
Cluster
Destination
Task
Assignment
System
layer
Heterogeneous
Spectrum
Self
grained
Mobility
Control
Sharing
Observations
Scalability
Stable
Correlated
Localization
Indoor
Determining
Optimization
calibrating
Experience
Construction
Connect


Machine Translation
Neural
Classification
Language
Model
Learning
Dependency
Parsing
Word
Segmentation
Character
Network
Text
Multi
Graph
Statistical
Machine
Method
Semantic
Parsing
Effective
Reading
Comprehension
Semantic
Role
Random Field
Role Labeling
Linguistic
Conditional Random
Selection
Segmentation
Tree
Feature
aware
Universal
Sequence
Phrase
Syntactic
Cross
Syntax
Masao Utiyama
Representation
Discourse Relation
Supervised
Recognition
Implicit Discourse
Dependency
Pinyin
Task
Information
Large scale
Utiyama Eiichro
End
An Empirical
Unsupervised
Tagging
Detection
Embedding
Efficient
Bilingual
Multilingual
Reordering
Framework
Extraction
Alignment
Study
Hierarchical
Continuous
Proceedings
Input
Conference
Algorithm
Pruning
Coreference
Eiichro Sumita
exploiting
Approach
Spam
Subword
Joint
Entailment
Named
NLP
Finding
Shallow
Comparison
Single
Spell
Supervised
Deep
BCMI
Global
Entropy
Aggregation
Association
Treebank
Maximum
driven
Converting
Natural
Constituent
Sentence
Deceptive
attention
Explicit
Corpus
Rule
System
Computational
Resolution
Encoder
Recurrent
Head
Convolutional
Opinion
Entity
Relation
Correction
Modeling
Mining
Labeling
augmented
Enhancement
International
Unified
Meta
Two
Dual
scale Hierarchical
Resource
Transformer
Recombination
Improving
Better
Uniform
Error
Improved
Penn
Open
Conversion
Labeled
Grammatical
Decoding
NTCIR
gram
Parser
Enhanced
Multiple
turn
Hybrid
Transition
as
into
Checking
Top
One
Accurate
Modular
Structure
Subsequence
Co

实体识别

简介

命名实体识别（Named Entity Recognition），简称为NER，是自然语言处理（NLP）中一项最基础的工作，它的任务就是识别出文本当中特定意义的实体。

MCU将其分为三大类：时间类（TIMEX)，实体类（EMAMEX）和数字类（NUMEX)，三大类又被分为七小类（Location, Person, Organization, Money, Percent, Date, Time），比如实践类包含人名，地名，机构名三类，时间类包含日期和时间两类，数字类包含货币和百分比两类。在知识图谱、情感分析、机器翻译、对话问答系统都有应用。

基本方法

(1) 基于规则和字典的方法

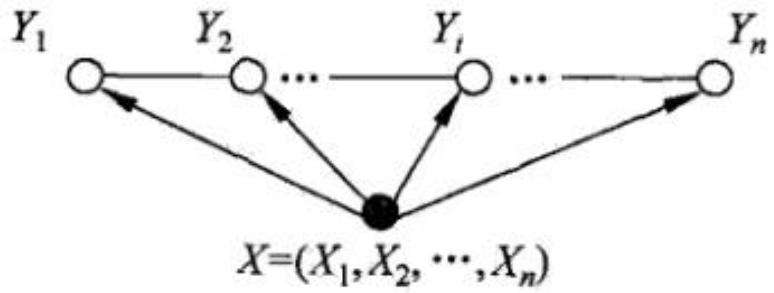
(2) 基于统计的传统的机器学习方法

隐马尔可夫模型（HMM)　　支持向量机（SVM)

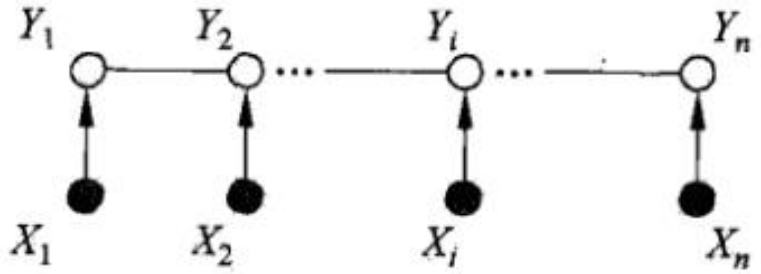
最大熵（ME)　　条件随机场（CRF)

(3) 基于深度学习的方法

条件随机场算法（CRF）



线性链条件随机场



X 和 Y 有相同的图结构的线性链条件随机场

用F(y,x)表示全局特征向量，即

$$F(y,x) = (f_1(y,x), f_2(y,x), \cdots, f_K(y,x))^{\mathrm{T}}$$

则条件随机场可以写成向量w和F(y,x)的内积的形式：

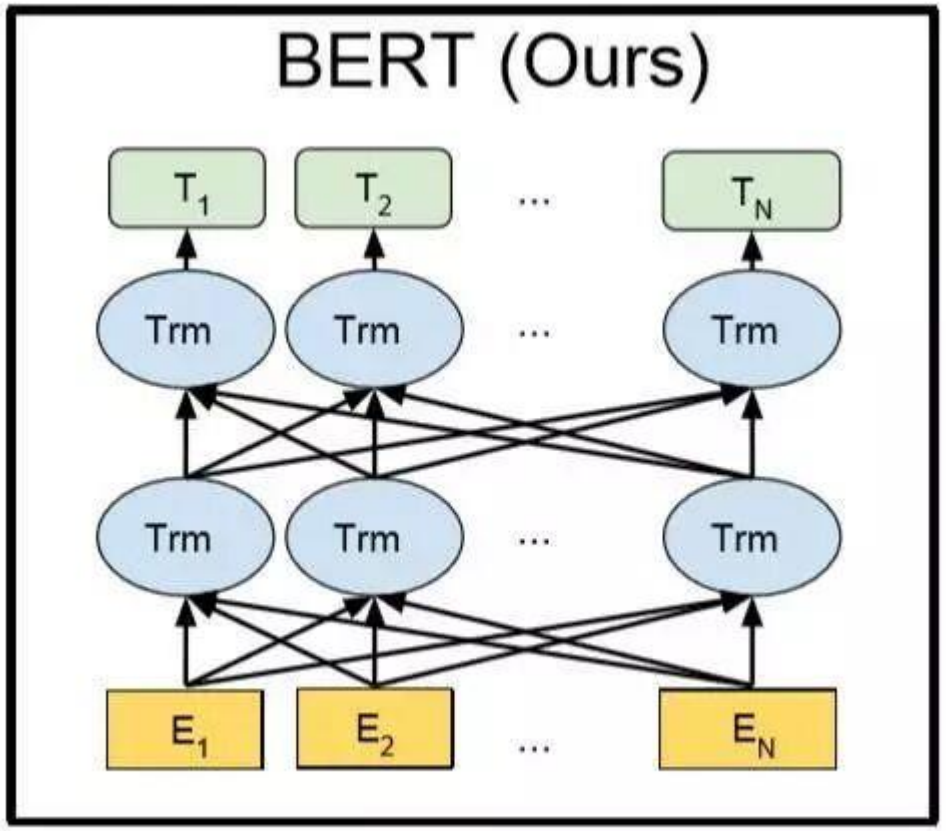
$$P_w(y|x) = \frac{\exp(w \cdot F(y,x))}{Z_w(x)}$$

其中

$$Z_w(x) = \sum_y \exp(w \cdot F(y,x))$$

BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding



Masked Language Model

Next Sentence Prediction

实体识别

SELECTED PUBLICATIONS

- Youjing Lu, **Fan Wu**, Shaojie Tang, Linghe Kong, and Guihai Chen, FREE: A Fast and Robust Key Extraction Mechanism via Inaudible Acoustic Signal, in Proceedings of the 20th ACM International Symposium on Mobile Ad Hoc Networking and Computing (**MobiHoc**), Catania, Italy, Jul. 2-5, 2019. (Corresponding Author)

bert ner--------------------

	Entity Name	Entity Type
0	Catania	LOC
1	Linghe Kong	PERSON
2	Youjing Lu	PERSON
2	Youjing Lu	PERSON
3	20th ACM International Symposium on Mobile Ad ...	MISC
4	Italy	LOC
5	Fan Wu	PERSON
6	Guihai Chen	PERSON
7	Shaojie Tang	PERSON
8	MobiHoc	MISC

stanford corenlp--------------------

	Entity Name	Entity Type
0	Author	TITLE
1	Linghe Kong	PERSON
2	Youjing Lu	PERSON
3	20th ACM International Symposium on Mobile Ad ...	MISC
4	Jul. 2 - 5	NUMBER
5	2019	DATE
6	Catania	CITY
7	Italy	COUNTRY
8	Guihai Chen	PERSON
9	Inaudible Acoustic Signal	ORGANIZATION
10	Shaojie Tang	PERSON
11	Wu	PERSON

Future Work

1、扩大数据集的范围，挖掘获取更多的关键信息

2、通过更完整的关键信息进一步分析学者之间的关系

EE447

感谢聆听

李辰轩　　林珈漫　　丁晨诗