Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

基于知识库的命名实体识别

硕士研究生　　尹继泽

2019年07月14日

背景简介

- 预期收获
  - 1. 了解自然语言处理中的一些基础知识
  - 2. 了解知识库在命名实体识别任务中的应用情况
  - 3. 理解如何基于知识库中的额外知识训练嵌入
  - 4. 了解引入知识库能够提升命名实体识别系统的效果

背景简介

- 算法提出的原因
  - 神经网络缓解了对特征工程的需求，但实践表明来自知识库的词典特征并不能被神经网络完全取代

基本概念

- 命名实体识别
  - 定义：从文本中识别命名实体的边界和类别
  - 比如：人名、地名、组织机构名、时间和数字表达（包括时间、日期、货币量和百分数等）

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

基本概念

- 知识库
  - 定义：一种使用计算机系统存储结构化和非结构化复杂信息的技术
    - 结构化数据是指由二维表结构来逻辑表达和实现的数据，严格地遵循数据格式与长度规范，主要通过关系型数据库进行存储和管理

id	name	gender	phone	address
1	张一	female	3337899	湖北省武汉市

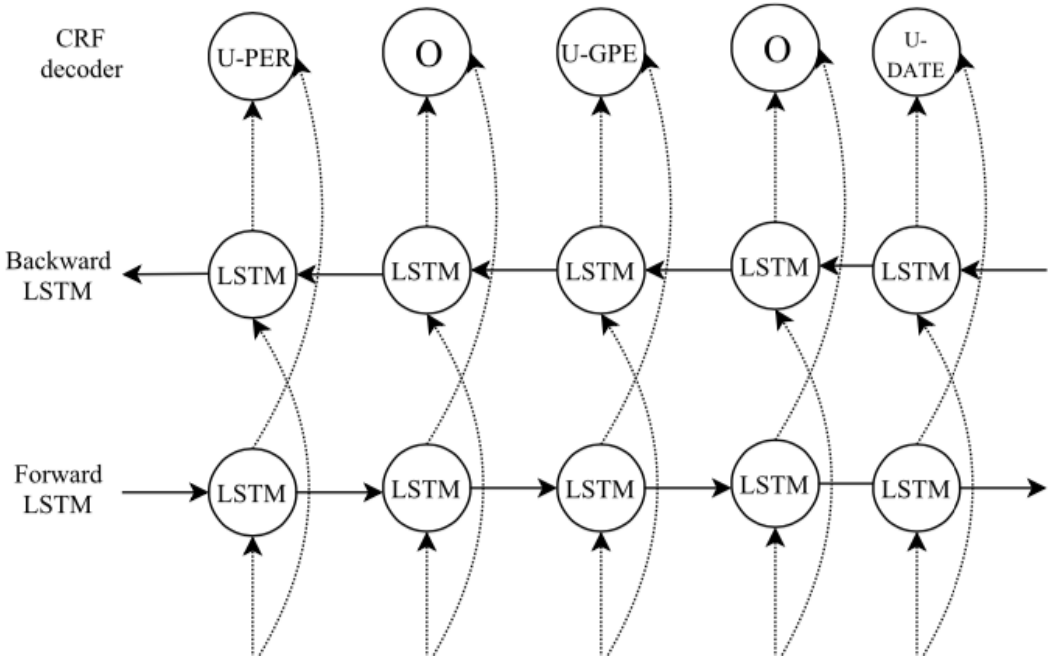
    - 非结构化数据是数据结构不规则或不完整，没有预定义的数据模型，不方便用数据库二维逻辑表来表现的数据（所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像和音频/视频信息等等）

基本概念

- 远程监督
  - 定义：将已有的知识库（比如 Wikipedia）对应到丰富的非结构化数据中（比如新闻文本），从而生成大量的训练数据，从而训练出一个效果不错的（命名实体识别）模型
  - 问题1：不完整——文本中还存在其他命名实体没有被标记出来
  - 问题2：噪声标记——被标记出来的命名实体是错误的

基本概念

- Bi-LSTM-CRF模型
  - 描述：一种流行的“神经网络+分类器” 模型
  - 参考：《基于LSTM-CRF的序列标注算法》、《长短期记忆网络（LSTM）》



算法原理

T	基于无标签数据和知识库，获得词典特征并结合神经网络提升命名实体识别效果
I	有标签语料，无标签语料，知识库（Wikipedia）【英文】
P	1. 获得词嵌入表示、字符嵌入表示、词模式特征向量、词典相似度向量 2. 构建神经网络-分类器模型 3. 训练模型
O	命名实体识别模型

P	神经网络缓解了对特征工程的需求，但实践中发现来自知识库的词典特征并不能被神经网络完全取代
C	存在有标签语料，无标签语料和知识库（Wikipedia）
D	如何获得词典特征的特征向量
L	CCF B类会议

算法原理

- 基于规则的知识库利用方法
  - 比如：正（反）向最大匹配
  - 问题：规则越来越复杂，有效期短，耗费人力和时间

举个栗子

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

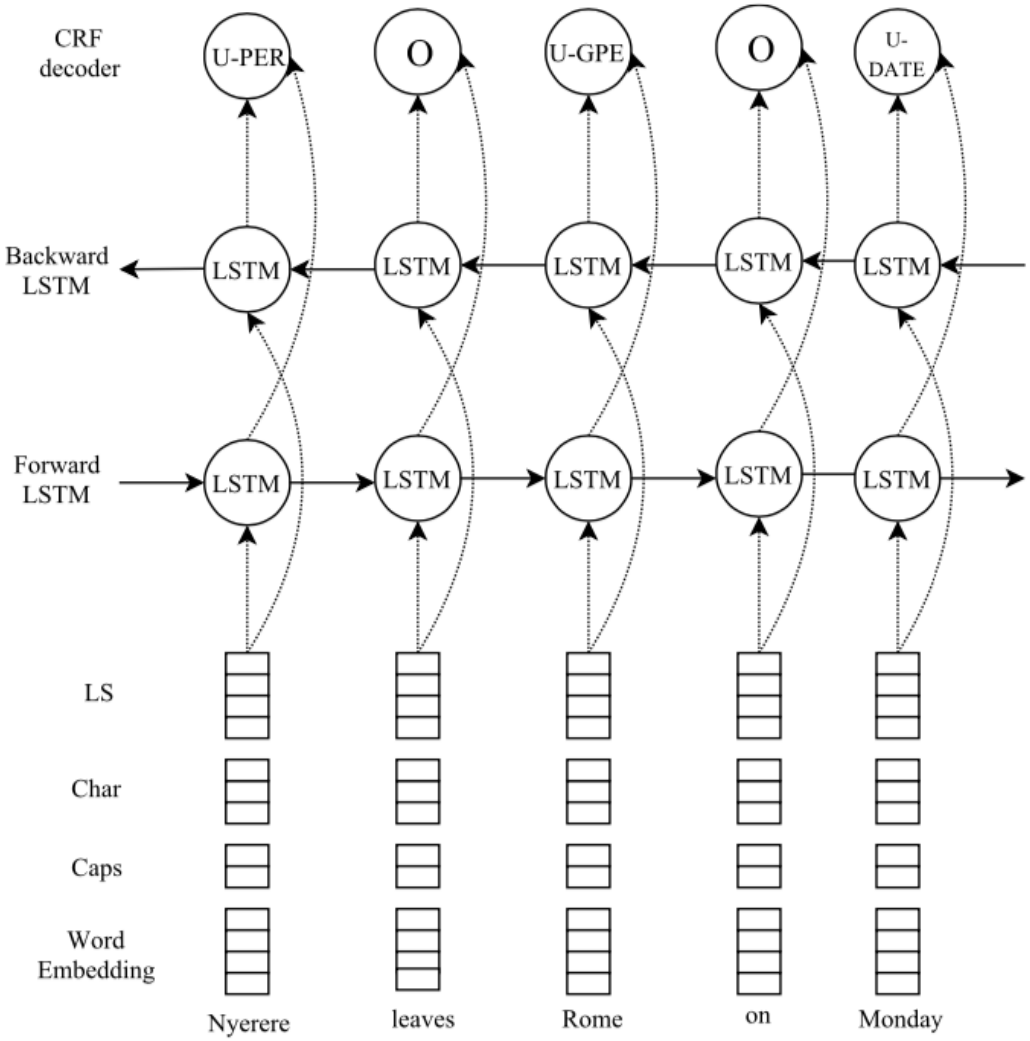
算法原理

- 基于特征工程的知识库利用方法
  - 比如：《A Study of the Importance of External Knowledge in the Named Entity Recognition Task》
  - 问题：特征来源于知识库而非目标领域语料，意味着特征不适合当前任务的可能性很大

Apple（某领域内统计的结果）		
公司	产品	水果
60%	30%	10%

算法原理

- 基于神经网络的知识库利用方法

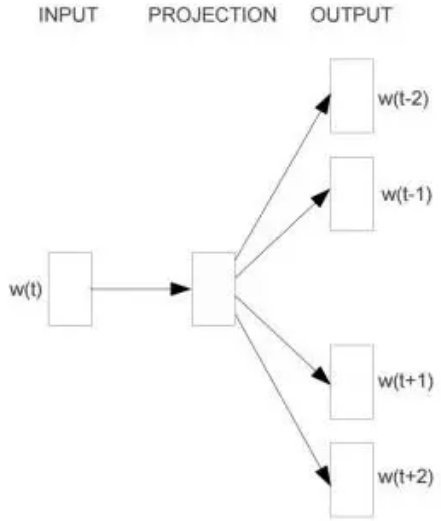


算法原理

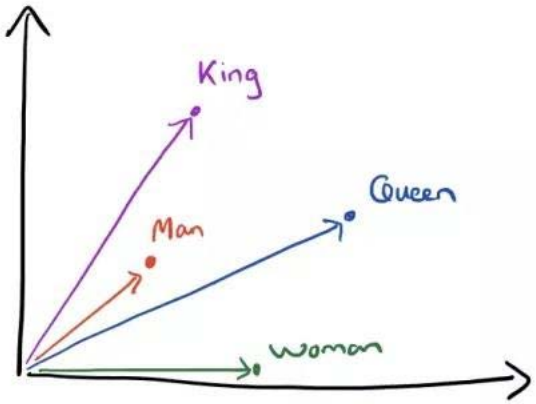
- 词嵌入表示
  - 定义：把一个维数为所有词的数量的高维空间嵌入到一个维数低得多的连续向量空间中，每个单词或词组被映射为实数域上的向量
    - Senna（Collobert et al., 2011）
    - Word2Vec（Mikolov et al., 2013）
    - GloVe（Pennington et al., 2014）
    - SSKIP（Yulia et al., 2015）【基于n-skip-gram模型】
  - 参考：《Deep Learning词向量生成——CBOW和Skip-gram》

算法原理

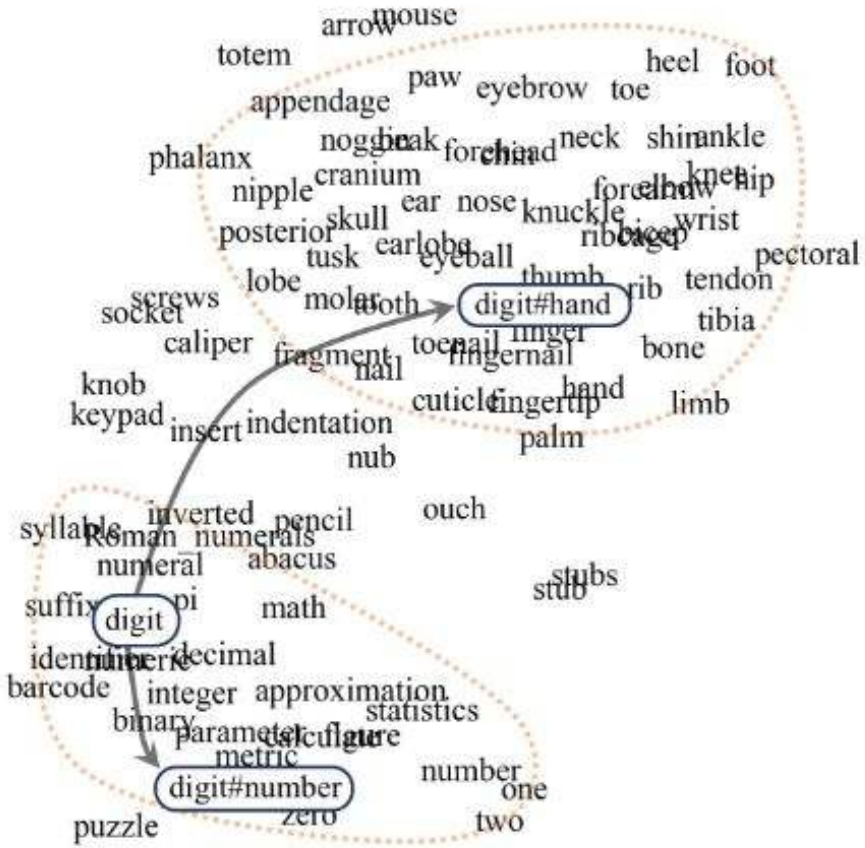
- 词嵌入表示



Skip-gram

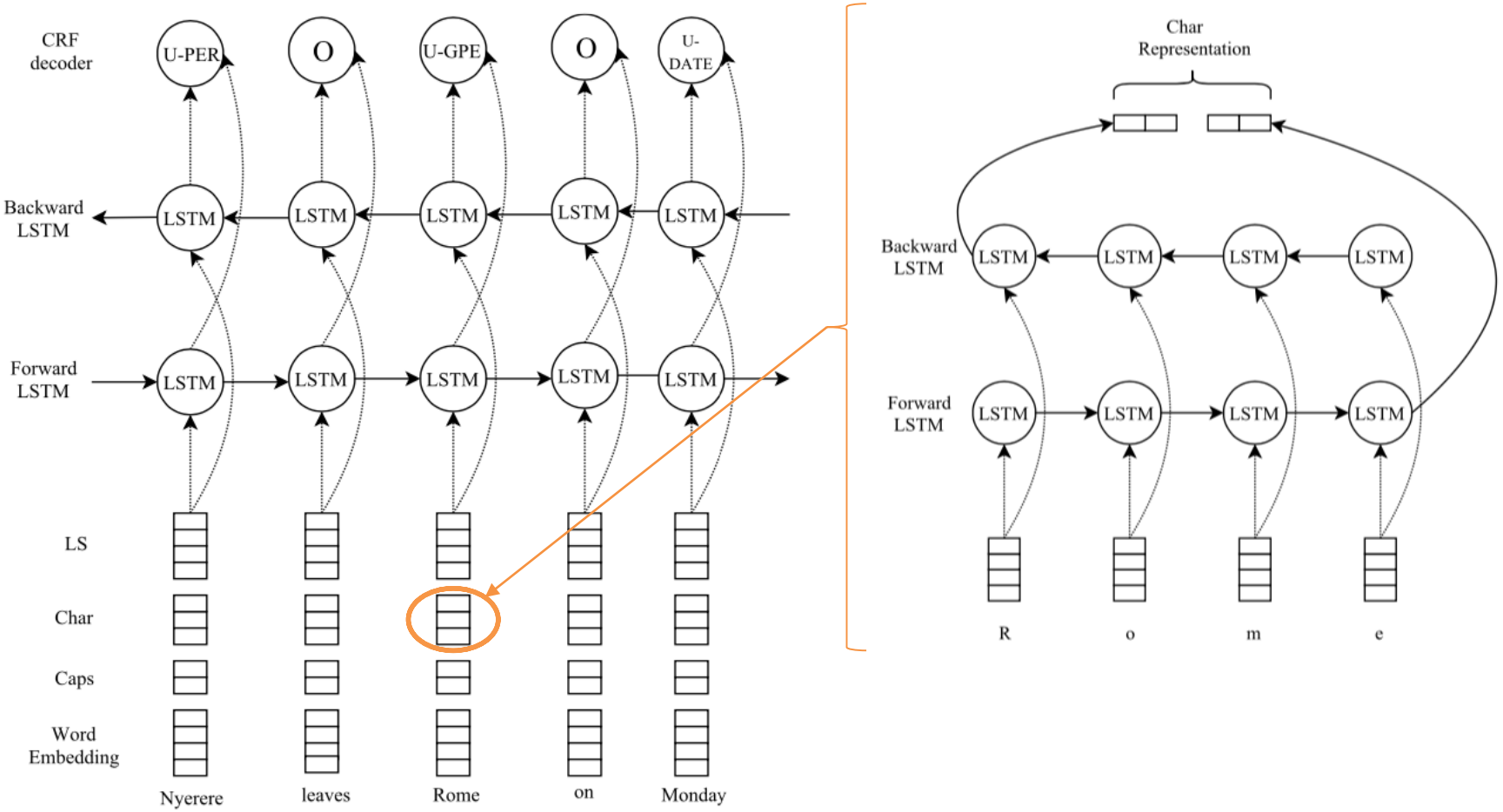


Word Vectors



算法原理

- 字符嵌入表示


CRF decoder
U-PER
O
U-GPE
O
U-DATE
Backward LSTM
Forward LSTM
LSTM
LS
Char
Caps
Word Embedding
Nyerere
leaves
Rome
on
Monday
Char Representation
R
o
m
e

算法原理

- 词模式特征向量
  - allUpper
  - allLower
  - upperFirst
  - upperNotFirst
  - numeric
  - noAlphaNum

算法原理

- 词典相似度向量
  - 基于Wikipedia，使用远程监督学习方法获得自动标记语料，包括3.2M的Wikipedia文章
  - 使用每一种实体类型的周围词来学习它对应的嵌入（skipgram模型），例如：为了得到软件类型的嵌入，需要基于所有被标记为软件的实体词周围的词来训练嵌入

(v1) On **October 9, 2009**, the **Norwegian Nobel Committee** announced that **Obama** had won the **2009 Nobel Peace Prize**.
(v2) On /date, the /organization/government_agency announced that /person/politician had won the /award.

算法原理

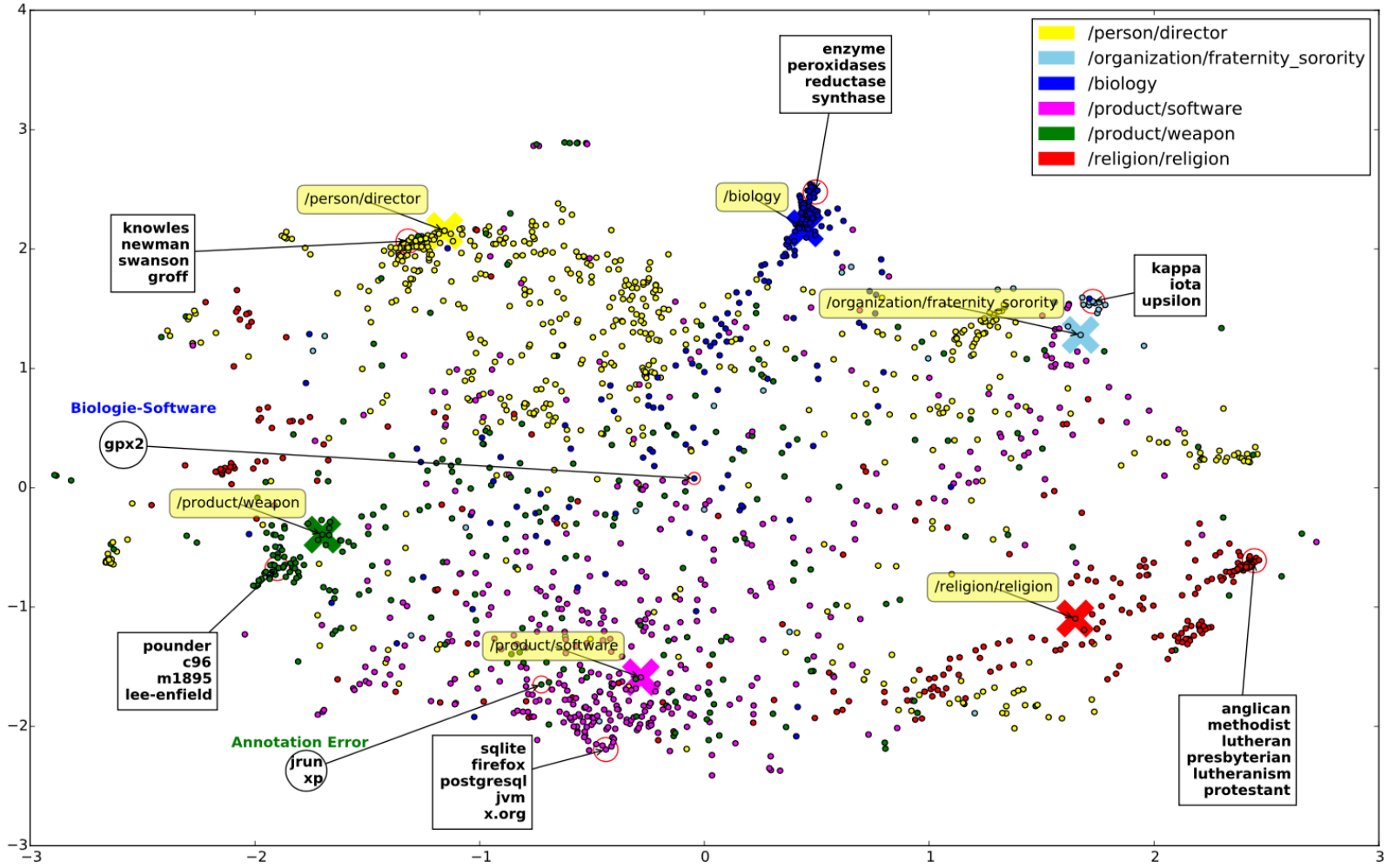


Figure 2: Two-dimensional representation of the vector space which embeds both words and entity types. Big Xs indicate entity types, while circles refer to words (i.e. named-entities, here).

算法原理

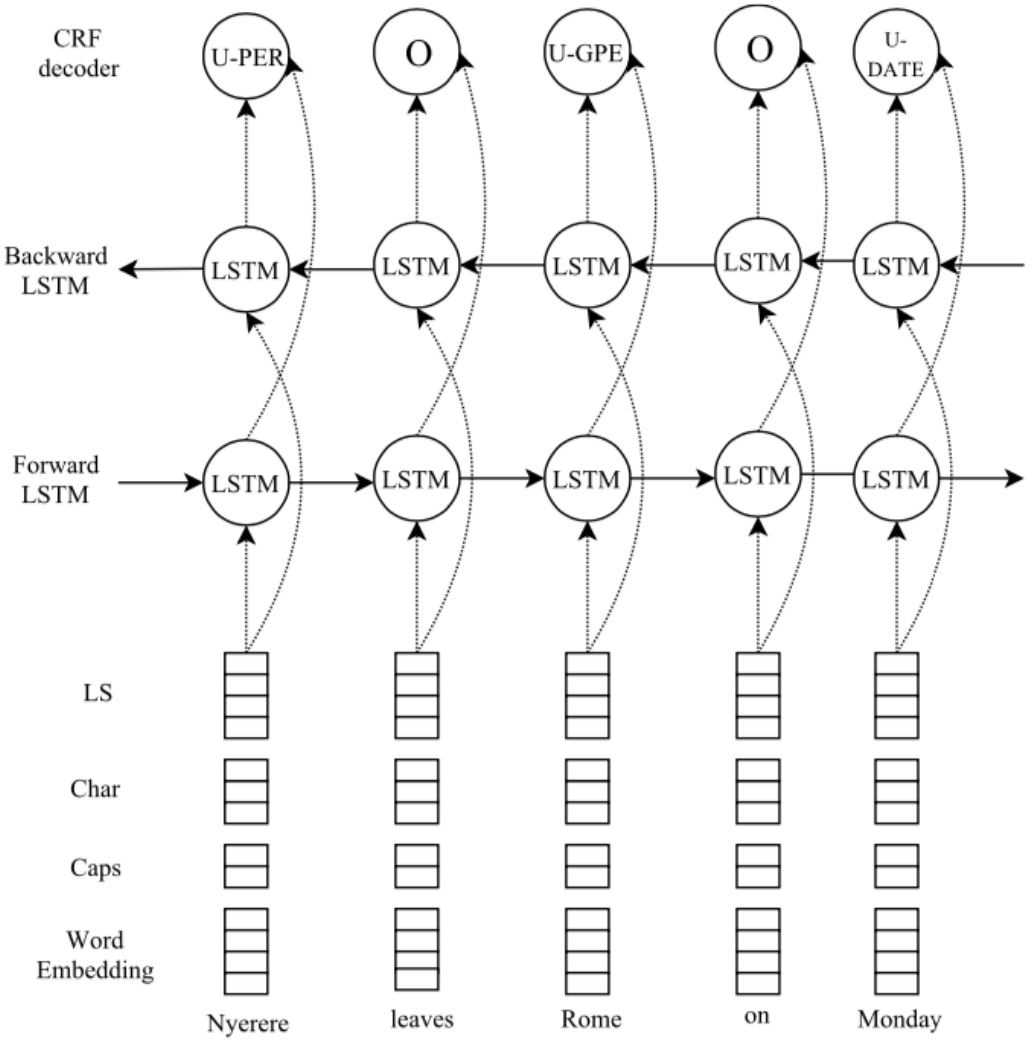
- 词典相似度向量
  - 得到数值向量形式的词典相似度表示
    - 维数为命名实体类型的数目
    - 每一维的大小等于词嵌入和当前维对应的命名实体类型嵌入的余弦相似度

Word	Entity Type	Sim	Word	Entity Type	Sim
hilton	/building/hotel	0.58	located	/location	0.47
	/building/restaurant	0.46		/location/city	0.44
	/person/actor	0.37		/building	0.40
gpx2	/biology	0.69	directed	/person/director	0.60
	/product/software	0.56		/art/film	0.55
jrun	/product/software	0.64	in	/date	0.58
	/product/weapon	0.23		/location/city	0.54
dammstadt	/location/city	0.45	won	/award	0.53
	/location/railway	0.44		/event/sports_event	0.53

Table 1: Topmost similar entity types to a few single-word mentions (left table) and non-entity words (right table).

算法原理

- 构建神经网络-分类器模型
- 训练模型



算法原理

- 语料统计信息

Dataset		Train	Dev	Test
CoNLL-2003	*#tok*	204,567	51,578	46,666
	#ent	23,499	5,942	5,648
OntoNotes 5.0	*#tok*	1,088,503	147,724	152,728
	#ent	81,828	11,066	11,257

Table 2: Statistics of the CoNLL-2003 and OntoNotes 5.0 datasets. *#tok* stands for the number of tokens, and *#ent* indicates the number of named-entities gold annotated.

算法原理

- 算法执行结果

Model	LEX	GAZ	CAP	EMB	CHE	LME	LS	F1
(Finkel et al., 2005)	+	+	+	•	•	•	•	86.86
(Ratinov and Roth, 2009)	+	+	+	•	•	•	•	90.88
(Lin and Wu, 2009)	+	+	+	•	•	•	•	90.90
(Luo et al., 2015)	+	+	+	•	•	•	•	91.20
(Collobert et al., 2011)	•	+	+	+	•	•	•	89.56
(Huang et al., 2015)	•	•	+	+	+	•	•	90.10
(Lample et al., 2016)	•	•	+	+	+	•	•	90.94
(Ma and Hovy, 2016)	•	•	+	+	+	•	•	91.21
(Shen et al., 2017)	•	•	+	+	•	•	•	90.89
(Strubell et al., 2017)	•	•	+	+	•	•	•	90.54
(Tran et al., 2017)	•	•	+	+	+	+	•	91.69
(Liu et al., 2017)	•	•	+	+	+	+	•	91.71
This work	•	+	+	+	+	•	+	**91.73**

Table 4: F1 scores on the CoNLL test set. The first four systems are feature-based, the others are neuronal. The feature configuration of each system is encoded with: LEX which stands for LEXical feature, GAZ for GAZetteers, CAP for CAPitalization, EMB for pre-trained EMBeddings, CHE for CHaracter Embeddings, LME for Language Model Embeddings, and LS for the proposed LS feature representation. + indicates that the model uses the feature set.

算法原理

- 算法执行结果

Model	LEX	GAZ	CAP	EMB	CHE	LME	LS	F1
(Finkel and Manning, 2009)	+	+	+	•	•	•	•	82.42
(Ratinov and Roth, 2009)	+	+	+	•	•	•	•	84.88
(Passos et al., 2014)	+	+	+	•	•	•	•	82.24
(Durrett and Klein, 2014)	+	+	+	•	•	•	•	84.04
(Chiu and Nichols, 2016)	•	+	+	+	+	•	•	86.28
(Shen et al., 2017)	•	•	+	+	+	•	•	86.52
(Strubell et al., 2017)	•	•	+	+	+	•	•	86.99
This work	•	+	+	+	+	•	+	**87.95**

Table 5: F1 scores on the ONTONOTES test set. The first four systems are feature-based, the following ones are neuronal. See Table 4 for an explanation of the column of features.

算法原理

- 算法执行结果

Model	BC	BN	MZ	NW	TC	WB
(Finkel and Manning, 2009)	78.66	87.29	82.45	85.50	67.27	72.56
(Durrett and Klein, 2014)	78.88	87.39	82.46	87.60	72.68	76.17
(Chiu and Nichols, 2016)	85.23	89.93	84.45	88.39	72.39	78.38
This work	**86.33**	**90.46**	**85.91**	**89.75**	**75.41**	**80.39**

Table 6: Per-genre F1 scores on ONTONOTES (numbers taken from Chiu and Nichols (2016)). BC = broadcast conversation, BN = broadcast news, MZ = magazine, NW = newswire, TC = telephone conversation, WB = blogs and newsgroups.

Model	CONLL	ONTONOTES
SSKIP	90.52 ($\pm$ 0.18)	86.57 ($\pm$ 0.10)
LS	89.94 ($\pm$ 0.16)	85.92 ($\pm$ 0.12)
all	**91.73 ($\pm$ 0.10)**	**87.95 ($\pm$ 0.13)**

Table 7: F1 scores of differently trained systems on CONLL and ONTONOTES 5.0 datasets. Capitalization (Section 4.2.3) and character features (Section 4.2.2) are used by default by all models.

优劣分析

- 优势
  - 词典相似度向量的计算与命名实体识别主要模型（Bi-LSTM-CRF模型）的训练和测试是独立的，因此不会在实际应用中增加计算负担
  - 既引入了词典特征，又结合了目标领域语料的上下文特征
- 劣势
  - 远程监督学习自动标注的语料存在不完整和噪声标记问题
  - 远程监督学习自动标注的语料受到知识库种类和规模的限制

应用总结

- 算法的应用领域
  - 前提：存在任务相关的数据和知识库
  - 目的：提升任务现有解决方案的效果
  - 思想：基于现有数据和知识库，确定映射f，输出包含数据信息和知识库中知识的数学表达，即

$$f(\text{数据信息}, \text{知识库中知识})$$

- 未来的发展
  - 针对社交媒体文本命名实体识别：存在集外词和低频词
  - 针对细粒度命名实体识别：缺乏人工标注的训练数据

参考文献

[1] Abbas Ghaddar, Philippe Langlais. Robust Lexical Features for Improved Neural Network Named-Entity Recognition[C]. Proceedings of the 27th International Conference on Computational Linguistics, 2018: 1896－1907.

[2] 宗成庆. 统计自然语言处理（第2版）[M]. 北京：清华大学出版社, 2013.

谢谢！

大成若缺，其用不弊。大盈若冲，其用不穷。大直若屈。大巧若拙。大辩若讷。静胜躁，寒胜热。清静为天下正。