给用户分享的文本自动“贴标签”——一个融合语义知识和阅读行为的深度学习模型

三人行语义沙龙, AI在图情—2019图书馆前沿技术论坛 (IT4L 2019),

8月13日, 上海图书馆

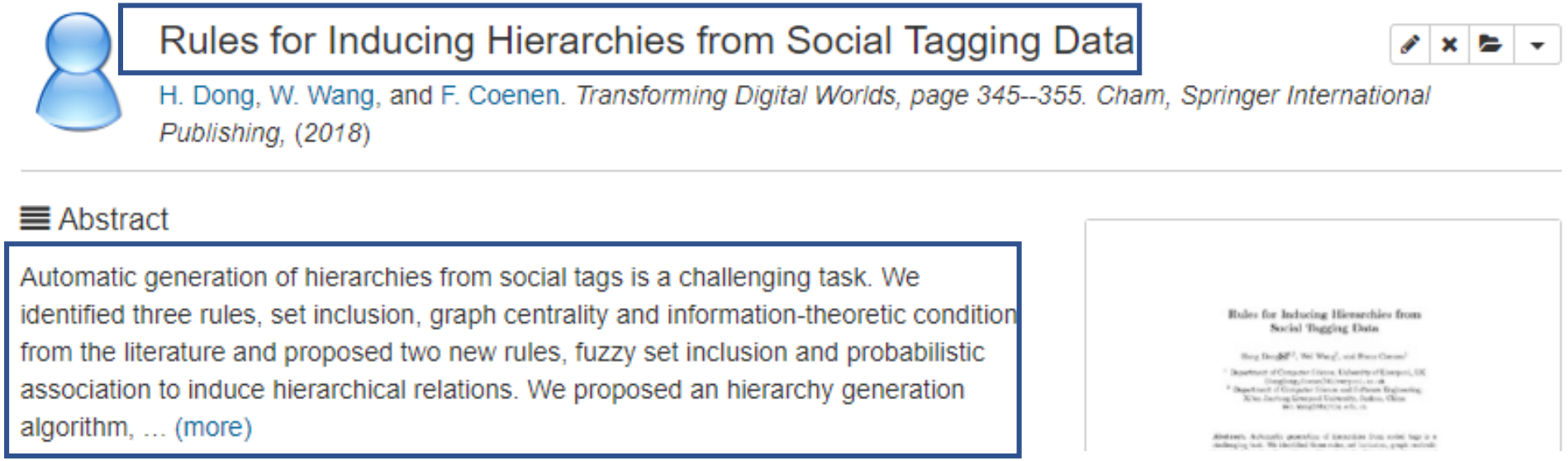
董行 (hang)

西交利物浦大学计算机科学系 博士生

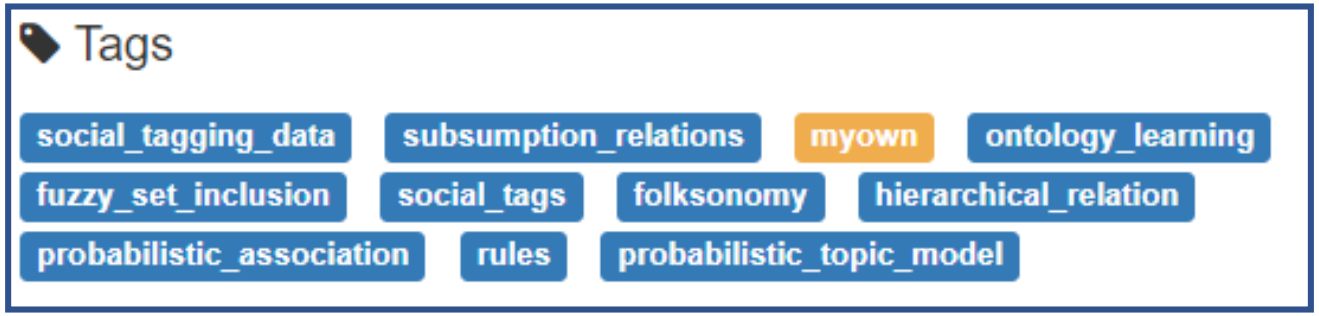
研究目的: 自动社会化标注

- 在社会化网站中，自动给用户分享的文档标注，以方便在线文本的组织、检索和推荐。（下图: Bibsonomy 网站中的社会化论文标注）

输入: 标题 + 摘要



输出: 标签



Images in website:
https://www.bibsonomy.org/bibtex/2e41f059d3b72c2bbfdb0063eaf8772b0

场景: 社会化问答系统中问题的自动标注

问题 (**标题** & 内容) + **标签层级关系** -> **标签**

克里斯蒂亚诺·罗纳尔多（Cristiano Ronaldo） 修改

克里斯蒂亚诺·罗纳尔多（Cristiano Ronaldo） » 话题组织 » 完整话题结构

父级话题 父话题是一个完全包括该话题的更大的话题。

克里斯蒂亚诺·罗纳尔多（Cristiano Ronaldo） 2018 年俄罗斯世界杯 葡萄牙国家男子足球队 摩洛哥国家男子足球队

「根话题」
- 产业
  - 职业
    - 体育从业者
      - 运动员
        - 足球运动员
          - 前锋
            - 克里斯蒂亚诺·罗纳尔多（Cristiano Ronaldo）

如何评价 2018 俄罗斯世界杯 C罗再破门，葡萄牙 1:0 摩洛哥?

世界杯-C罗4分钟破门葡萄牙胜 摩洛哥遭连败出局

相关问题C罗622球成为进球最多葡萄牙人，是否说明了C罗的历史地位超过了尤西比奥?

Images in website:
[1] https://biendata.com/competition/zhihu/
[2] https://www.zhihu.com/question/281774641/answer/422318818
[3] https://www.zhihu.com/topic/19577949/organize/entire#anchor-children-topic

微博自动标注

自动给用户发布的微博添加标签，以方便微博的检索与管理。

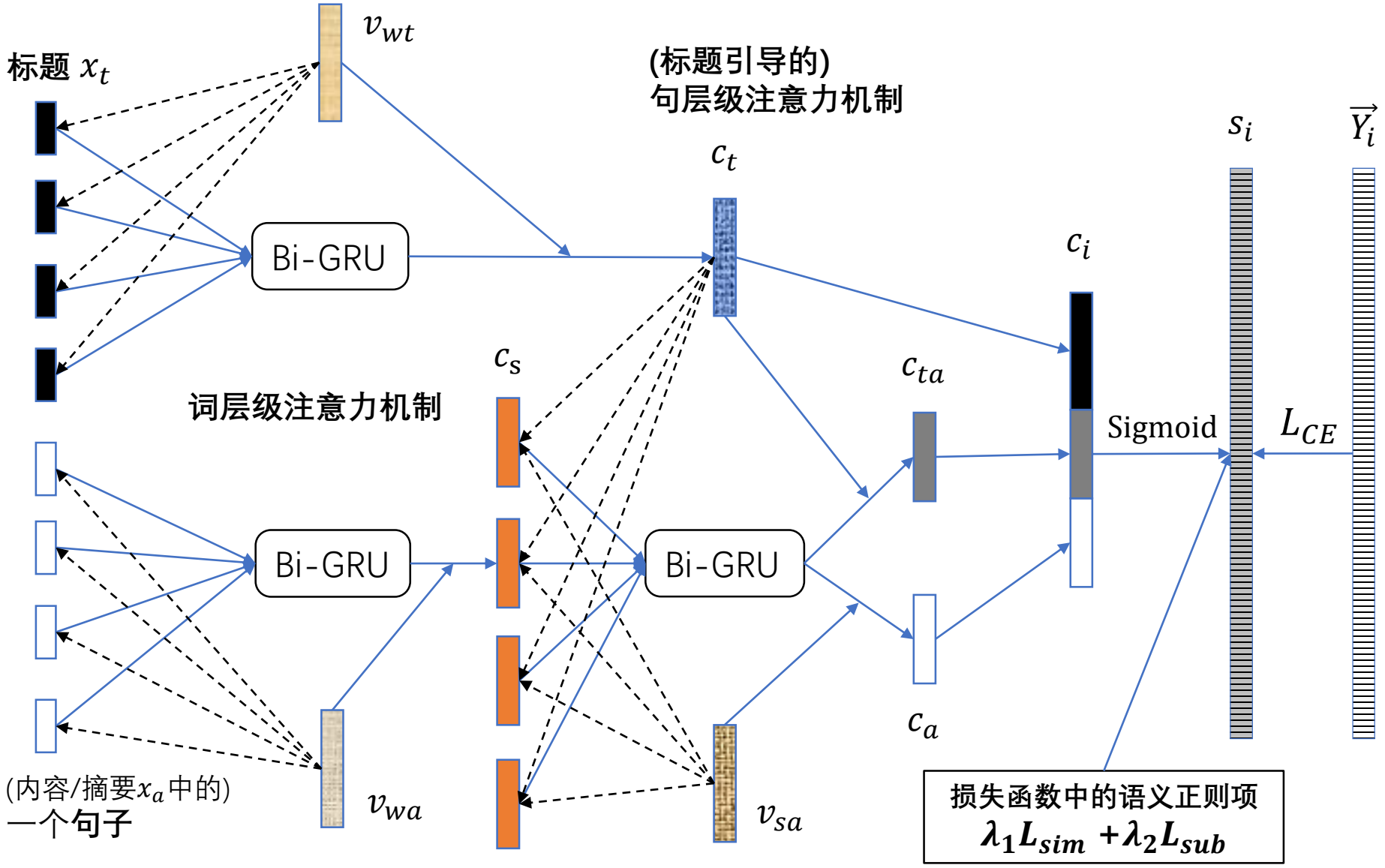
Images from: [1] https://weibo.com/2993049293/Gmlue2QLO?type=comment
[2] https://twitter.com/OptaJoe/status/1009407388246577152

当前的方法与问题

- 多标签分类 (Multi-label classification)，须要考虑标签间的关系。
- 方法: 注意力机制的神经网络 (Bi-GRU (Cho *et al.*, 2014), HAN (Yang *et al.*, 2016))

- 问题: (1) 在预测标签时没有融入标签关系（知识组织系统 KOS）
  - 相似关系
  - 层级关系

  (2) 在“阅读”文档时，忽视了元数据层次之间的关系
  - 题名对于阅读和标注的重要性: 具有很强的表达性和区分性 (Figueiredo *et al.*, 2013)
  - 题名 与 摘要 (内容) 之间的关系



Coloured version of the Figure in H. Dong, W. Wang, K. Huang, F. Coenen, Joint Multi-Label Attention Networks for Social Text Annotation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HLT 2019), Volume 1 (Long and Short Papers), pp. 1348-1354.

损失函数中的语义正则项 (semantic-based loss regularisers)

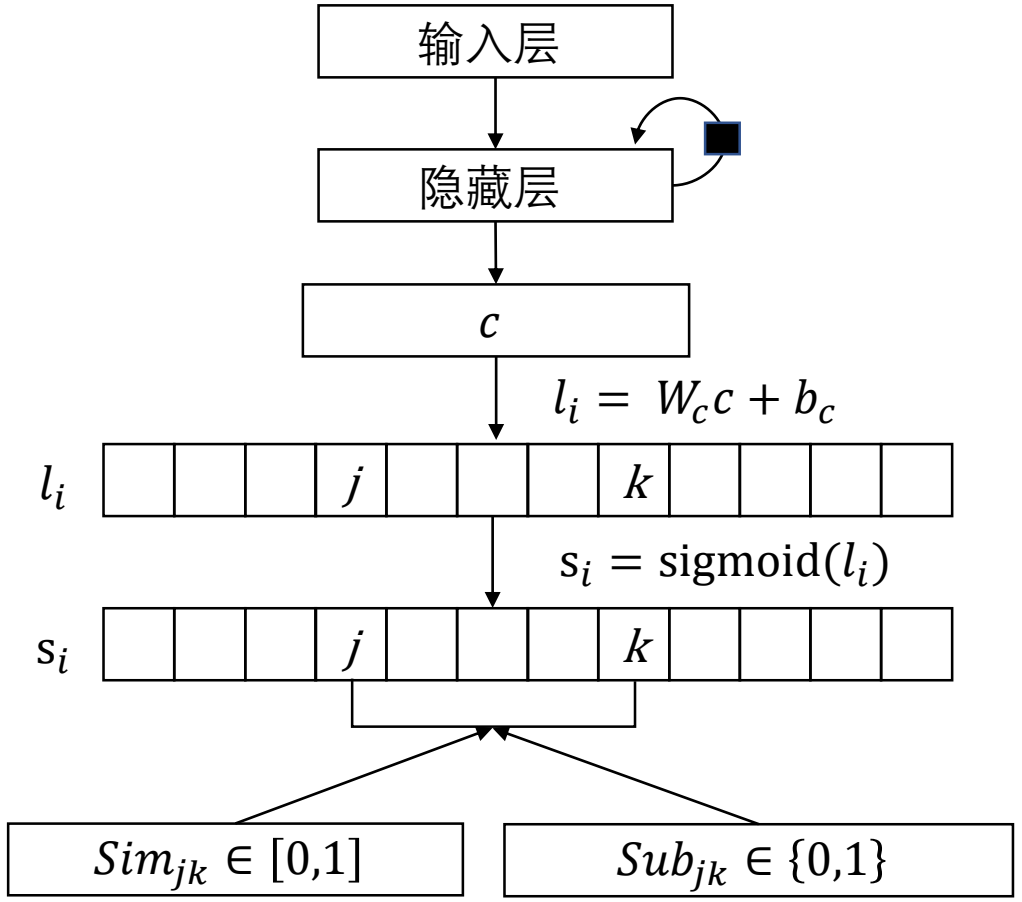
$$L = L_{CE} + \lambda_1 L_{sim} + \lambda_2 L_{sub}$$

$$L_{CE} = -\sum_i \sum_j (y_{ij} \log(s_{ij}) + (1 - y_{ij}) \log(1 - s_{ij}))$$

$$L_{sim} = \frac{1}{2} \sum_i \sum_{j,k|y_j,y_k \in Y_i} Sim_{jk} |s_{ij} - s_{ik}|^2$$

$$L_{sub} = \frac{1}{2} \sum_i \sum_{j,k|y_j,y_k \in Y_i} Sub_{jk} R(s_{ij})(1 - R(s_{ik}))$$

- $Sim_{jk} \in [0,1]$ 表示 第j个标签 和 第k个标签 的相似程度.
- $Sub_{jk} \in \{0,1\}$ 表示 第j个标签 是否是 第k个标签 的子标签.



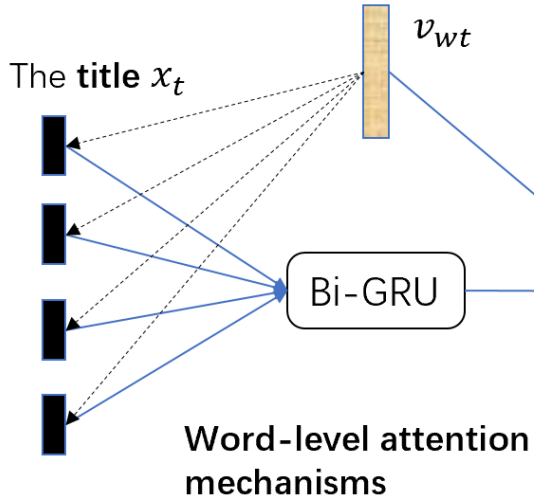
标题引导的注意力机制

- 词层级注意力机制

$$v^{(i)} = \tanh(W_t h^{(i)} + b_t)$$

$$\alpha^{(i)} = \frac{\exp(v_{wt} \bullet v^{(i)})}{\sum_{i \in [1, n_t]} \exp(v_{wt} \bullet v^{(i)})}$$

$$c_a = \sum_{i \in [1, n_t]} \alpha^{(i)} h^{(i)}$$

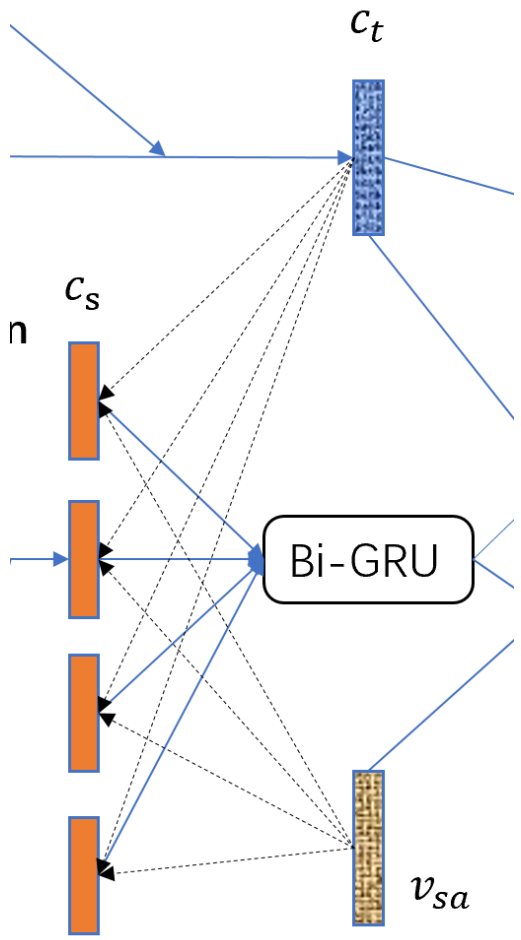


- 标题引导的句层级注意力机制

$$v_s^{(r)} = \tanh(W_s h_s^{(r)} + b_s)$$

$$\alpha_s^{(r)} = \frac{\exp(c_t \bullet v_s^{(r)})}{\sum_{k \in [1, n_s]} \exp(c_t \bullet v_s^{(k)})}$$

$$c_{ta} = \sum_{r \in [1, n_s]} \alpha_s^{(r)} h_s^{(r)}$$



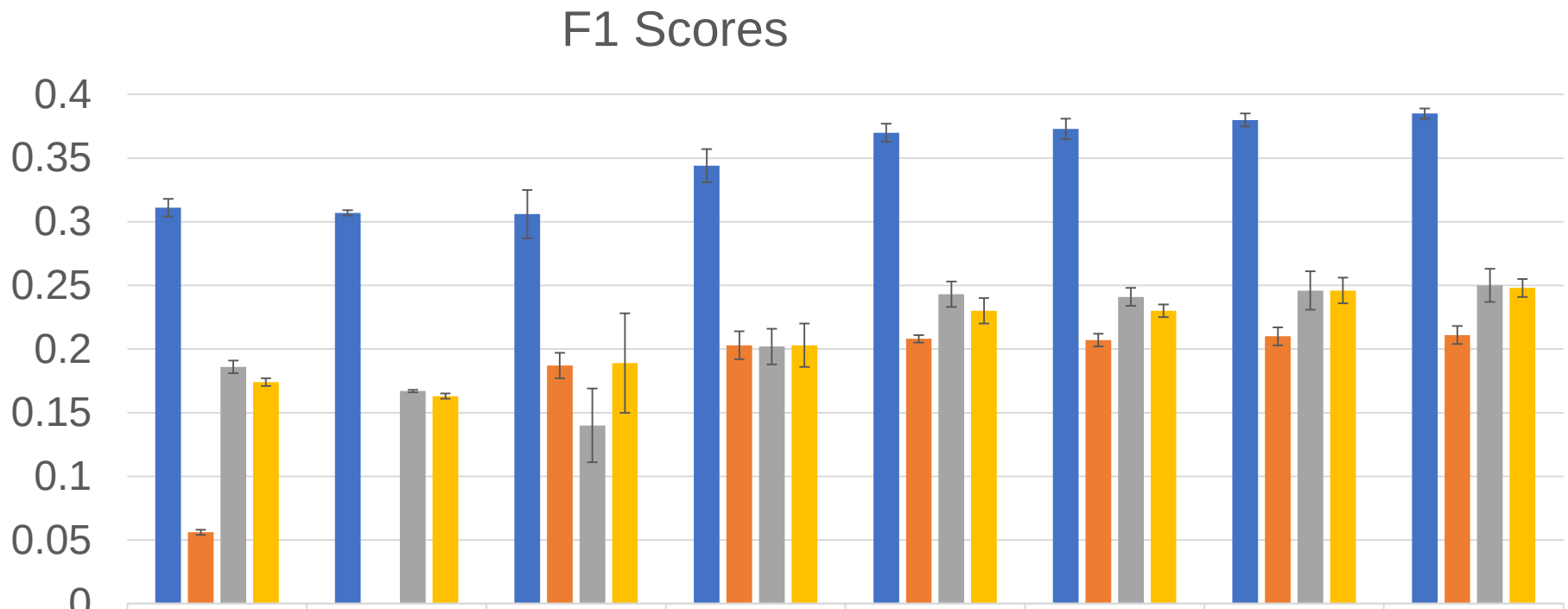
实验设置: 数据集和实现

- 社会化论文标注 (Bibsonomy, CiteULike) 和问题标注 (知乎 Zhihu)
- 数据集统计: $|X|$, 文档数; $|Y|$, 标签数; $|V|$, 词汇数; Ave, 文档对应的平均标签数; $\sum Sub$, 标签层级关系对的数量

Dataset	$\|X\|$	$\|Y\|$	$\|V\|$	Ave	Σ_{Sub}
Bibsonomy (clean)	12,101	5,196	17,619	11.59	101,084
CiteULike-a (clean)	13,319	3,201	17,489	11.60	107,273
CiteULike-t (clean)	24,042	3,528	23,408	7.68	141,093
Zhihu (sample)	108,168	1,999	62,519	2.45	2,655

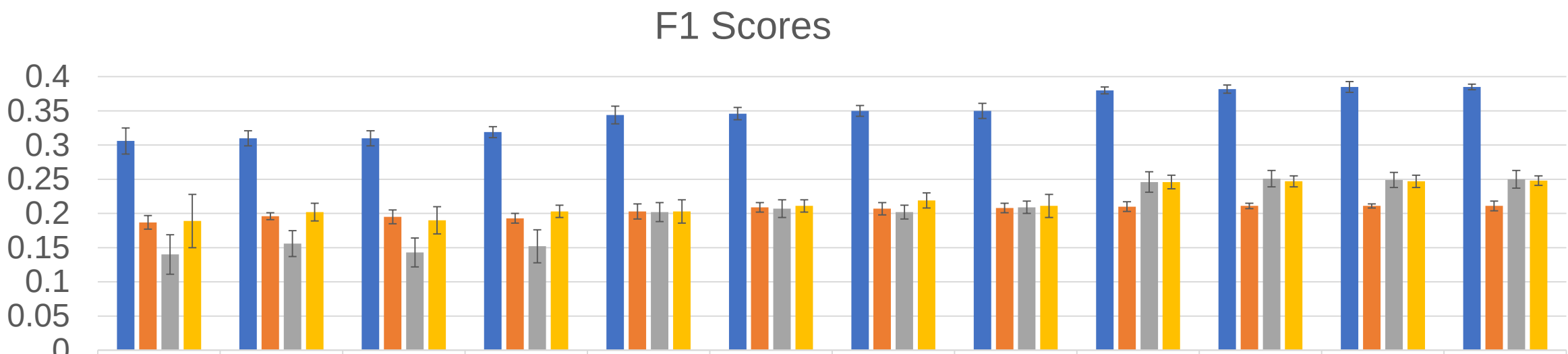
- 20% 作为最终测试数据, 80% 进行10折交叉验证
- 标签相似矩阵Sim_{jk}通过计算标签集嵌入embedding的余弦距离获得;
- 标签层级矩阵Sub_{jk}通过匹配Microsoft Concept Graph获得 (知乎除外，知乎提供用户编辑的标签层级关系)
- 参数: 词嵌入维度100; 隐藏层节点数100; 最大文本长度 对于Bibonomy, CiteULike 300个词 (句长30）, 知乎 100个词 （句长25） …

结果 (1) – 对比基准方法



	LDA	SVM-rbf	Bi-GRU	HAN	JMAN-s-tg	JMAN-s-att	JMAN-s	JMAN
Bibsonomy-F1	0.311	0.307	0.306	0.344	0.37	0.373	0.38	0.385
Zhihu-F1	0.056		0.187	0.203	0.208	0.207	0.21	0.211
CiteULike-a-F1	0.186	0.167	0.14	0.202	0.243	0.241	0.246	0.25
CiteULike-t-F1	0.174	0.163	0.189	0.203	0.23	0.23	0.246	0.248

结果 (2) – 关于语义正则项



	Bi-GRU	Bi-GRU+sim	Bi-GRU+sub	Bi-GRU+simsub	HAN	HAN+sim	HAN+sub	HAN+simsub	JMAN-s	JMAN+sim	JMAN+sub	JMAN
Bib-F1	0.306	0.31	0.31	0.319	0.344	0.346	0.35	0.35	0.38	0.382	0.385	0.385
Zhihu-F1	0.187	0.196	0.195	0.193	0.203	0.209	0.207	0.208	0.21	0.211	0.211	0.211
C-a-F1	0.14	0.156	0.143	0.152	0.202	0.207	0.202	0.209	0.246	0.251	0.249	0.25
C-t-F1	0.189	0.202	0.19	0.203	0.203	0.211	0.219	0.211	0.246	0.247	0.247	0.248

结果 (3) – 模型运行和收敛速度

Comparison of training time for all models in seconds

	SVM	LDA	Bi-GRU	Bi-GRU+s	HAN	HAN+s	JMAN-s-tg	JMAN-s-att	JMAN-s	JMAN
Bib	1107 ± 12	**110 ± 2**(1)	1480 ± 92	1683 ± 78	1164 ± 52	1434 ± 74	1075 ± 87	**1024 ± 100**(3)	**894 ± 55**(2)	1138 ± 86
C-a	1660 ± 31	**113 ± 3**(1)	869 ± 288	877 ± 57	462 ± 63	554 ± 45	434 ± 49	**429 ± 41**(3)	**394 ± 33**(2)	468 ± 38
C-t	4796 ± 50	**210 ± 7**(1)	1635 ± 1034	1469 ± 276	858 ± 100	947 ± 115	**752 ± 52**(3)	780 ± 69	**744 ± 62**(2)	839 ± 49
Zhi	over 1 day	**903 ± 31**(1)	1455 ± 69	2459 ± 151	1387 ± 78	2388 ± 275	**1220 ± 81**(3)	1275 ± 99	**1147 ± 44**(2)	1712 ± 105

Training time of the three most efficient models are in **bold** and marked with a ranking index in brackets. BiGRU+s and HAN+s denote the models with semantic-based loss regularisers.

(a) Bibsonomy (b) CiteUlike-a (c) CiteUlike-t (d) Zhihu

Fig. 3. Convergence plot: training loss with respect to the number of training epochs for the Bi-GRU, HAN, JMAN-s and JMAN models

可视化: 注意力机制和标注结果 (Bibsonomy的测试数据)

ori	tg			title:	an	information	visualization	tool	for	personalized
		knowledge	work	in	many	fields	requires	examining	several	aspects
		attain	meaningful	understanding	that	is	not	explicitly	available	
		despite	recent	advances	in	document	corpus	visualization	research	,
		principled	approaches	which	enable	the	users	to	personalize	the
		in	this	paper	,	we	present	information	visualization	for
		innovative	visualization	tool	which	employs	the	personal	information	model
		not	only	does	the	tool	allow	the	users	to
		exploration	and	analysis	of	a	document	collection	,	it
		the	usability	of	the	tool	was	evaluated	and	the
		it	worthwhile	to	conduct	a	usability	study		

prediction:	user	information	visualization	information_visualization				
labels:	user	information	interface	user_interface	semantic	social	management	
	ontology	visualization	personal	information_visualization	exploratory	semantic_desktop	desktop	personal_inform

紫色块显示句子（占标题title下的每两行）中每个词对于标注的重要性;

红色块显示 “ori” (原始的) 和 “tg” (标题引导的) 句层级注意力权重，表示文档中每个句子对于标注的重要性。颜色越深则权重越大。

文档下方展示了标注结果 (prediction) 和真实的用户生成的标签 (labels)。

更多结果可参见 https://github.com/acadTags/Automated-Social-Annotation/tree/master/0%20JMAN/results

可视化: 注意力机制和标注结果 (Bibsonomy的测试数据)

ori	tg										
			title:	chinese	culture	and	ecommerce	an	exploratory	study	
		differing	characteristics	of	local	environments	,	both	infrastructural	and	socioeconomic
0.02	0	,	have	created	a	significant	level	of	variation	in	the
		acceptance	and	growth	of	ecommerce	in	different	regions	of	the
0.16	0.01	this	paper	focuses	on	the	impact	of	these	infrastructural	and
		socioeconomic	factors	on	ecommerce	development	in	china			
		the	findings	provide	insights	into	the	role	of	culture	in
0.16	0.11	ecommerce	,	and	the	factors	that	may	impact	a	broader
		acceptance	and	development	of	ecommerce	in	china			
		in	this	paper	,	we	present	and	discuss	our	findings
0.16	0.12	,	and	identify	changes	that	will	be	required	for	broader
		acceptance	and	diffusion	of	ecommerce	in	china			
		cultural	issues	such	as	socializing	effect	of	commerce	,	transactional
0.31	0.38	and	institutional	trust	,	and	attitudes	toward	debt	were	determined
		to	be	the	major	impediments	to	ecommerce	in	china	
		however	,	our	research	also	shows	that	,	even	though
0.16	0.36	their	means	for	payment	are	different	,	the	most	enlightened
		,	able	,	and	sophisticated	consumers	in	china	participate	in

prediction:	**culture**	**e_commerce**	**china**	**chinese**				
labels:	study	culture	e_commerce	china	office	commerce	intercultural_communication	chinese

可视化: 注意力机制和标注结果 (CiteULike-t的测试数据)

ori	tg				**title:**	virtual	machines	versatile	platforms	for	systems	
		virtual	machine	technology	applies	the	concept	of	virtualization	to	an	e
		machine	compatibility	constraints	and	hardware	resource	constraints	to	enable	a	hi
		virtual	machines	are	rapidly	becoming	an	essential	element	in	computer	sy
		they	provide	system	security	,	flexibility	,	cross	platform	compatibility	
		efficiency										
		designed	to	solve	problems	in	combining	and	using	major	computer	sy
		technologies	play	a	key	role	in	many	disciplines	,	including	ope
		for	example	,	at	the	process	level	,	virtualizing	technologies	su
		platform	independent	network	computing							
		at	the	system	level	,	they	support	multiple	operating	system	enviro
		platform	and	in	servers							
		br	br	historically	,	individual	virtual	machine	techniques	have	been	dev
		that	employ	them	in	some	cases	they	even	referred	to	
		in	this	text	,	smith	and	take	a	new	approach	
		a	unified	discipline								
		pulling	together	cross	cutting	technologies	allows	virtual	machine	implementations	to	
		a	well	structured	manner							
		topics	include	instruction	set	emulation	,	dynamic	program	translation	and	optin
		machines	including	java	and	,	and	system	virtual	machines	for	b

prediction:	**machine**	**virtual**	comp	**virtualization**	virtual_machine
labels:	book	machine	virtual	text_book	virtualization

结论

- 现有的神经网络可以从融合知识和用户行为的角度提升性能。在自动社会化标注中，这主要体现在
  - 融合知识: 用标签间相似和层级关系约束多标签分类的输出层。
  - 用户行为: 在阅读和标注时，从题名字段出发，然后引导句子的理解，会产生更好的标注效果。
- 未来的研究: 探索融合知识和用户行为的新方法，提升神经网络在相关应用上的性能与可解释性。
  - 不同的知识呈现形式
  - 用户行为的模式
  - 新的模型结构和融合方式

参考文献和推荐阅读

[1] Dong, H., Wang, W., Huang, K., & Coenen, F. (2019, June). Joint Multi-Label Attention Networks for Social Text Annotation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 1348-1354). （自动社会标注的深度学习模型）

关于深度学习与注意力机制:

[2] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, Conference.* http://arxiv.org/abs/1409.0473 (注意力机制最早在机器翻译中的提出和应用)

[3] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016, June). Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 1480-1489). （层级注意力网络 HAN 在文本分类中的提出和应用）

[4] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014, October). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1724-1734). (双向门控循环单元 Bi-GRU 在机器翻译中的提出和应用)

关于自动社会化标注:

[4] Figueiredo, F., Pinto, H., Belém, F., Almeida, J., Gonçalves, M., Fernandes, D., & Moura, E. (2013). Assessing the quality of textual features in social media. *Information Processing & Management*, *49*(1), 222-247. （不同元数据字段作为文本特征的质量研究：其中一个结论是文本的题名字段具有很强的表达性和区分性）

[5] Hassan, H. A. M., Sansonetti, G., Gasparetti, F., & Micarelli, A. (2018, September). Semantic-based tag recommendation in scientific bookmarking systems. In *Proceedings of the 12th ACM Conference on Recommender Systems* (pp. 465-469). ACM. (社会化科技文献标注)

[6] Nie, L., Zhao, Y. L., Wang, X., Shen, J., & Chua, T. S. (2014). Learning to recommend descriptive tags for questions in social forums. *ACM Transactions on Information Systems (TOIS)*, *32*(1), 5. （社会化问答系统中的问题标注）

谢谢聆听

董行 | hangdong@liverpool.ac.uk

代码开源: https://github.com/acadTags/Automated-Social-Annotation
海报: https://www.aclweb.org/anthology/attachments/N19-1136.Poster.pdf
主页: http://cgi.csc.liv.ac.uk/~hang/index.html