

《第十七届中国计算语言学大会》 学生研讨会
CCL 2018, 湖南, 长沙



如何端到端地写科研论文?

邱锡鹏

复旦大学

2018年10月23日

<http://nlp.fudan.edu.cn/xpqi>

为什么要写科研论文 (Scientific Research Paper) ?



- ▶ 提出一种新的**问题、观点、方法或实验手段**等。
 - ▶ 有价值 (理论价值和应用价值)
 - ▶ 负面结果 (Negative Results) \neq 没价值
- ▶ 具有共享精神, 希望为一个领域做出贡献
- ▶ 树立自己在某个领域的学术地位

- ▶ 其它动机
 - ▶ 职称压力 (对于大多数教师)
 - ▶ 毕业要求 (对于大多数学生)



什么是一个好的科研论文?

▶ 好的论文=好的研究+好的写作

▶ 好的研究 比写作技巧更重要，一个另外的话题

▶ 良好定义 (well-defined) 的问题

▶ “一流高手提问题、二流高手解问题、三流高手抄问题”

▶ 好的写作

▶ 以读者为中心

▶ 逻辑清晰

▶ 可重复实现



为什么发表论文很难？

- ▶ 好的期刊会议录用率都很低 ~20%
- ▶ 不同期刊、会议的论文风格不同
- ▶ 没有严格的标准，很难评价
 - ▶ 同行评审 (peer-review)
 - ▶ 一把双刃剑
 - ▶ 拒绝一篇论文比录用一篇论文风险更低
- ▶ 非“面对面”交流

Hinton, G. E. and Nowlan, S. J.

How learning can guide evolution.

<http://www.cs.toronto.edu/~hinton/papers.html>

Complex Systems, 1, 495--502. [[pdf](#)]

This paper was rejected by the Cognitive Science Society Conference (against the advice of the referees) because the chairman of the organizing committee thought it might mislead cognitive scientists.

September 29, 1955

被拒的诺奖论文

▶ Rosalyn Yalow

- ▶ The invention of the radioimmunoassay (放射免疫分析法)

- ▶ Nobel Prize in Physiology and Medicine in 1977

▶ More

- ▶ <https://www.sciencealert.com/these-8-papers-were-rejected-before-going-on-to-win-the-nobel-prize>

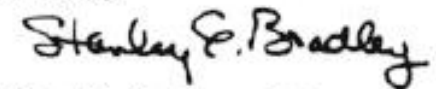
Dr. Solomon A. Berson
Radioisotope Service
Veterans Administration Hospital
130 West Kingsbridge Road
Bronx 68, New York

Dear Dr. Berson:

I regret that the revision of your paper entitled "Insulin- I^{131} Metabolism in Human Subjects: Demonstration of Insulin Transporting Antibody in the Circulation of Insulin Treated Subjects" is not acceptable for publication in THE JOURNAL OF CLINICAL INVESTIGATION. -----

----- The second major criticism relates to the dogmatic conclusions set forth which are not warranted by the data. The experts in this field have been particularly emphatic in rejecting your positive statement that the "conclusion that the globulin responsible for insulin binding is an acquired antibody appears to be inescapable". They believe that you have not demonstrated an antigen-antibody reaction on the basis of adequate criteria, nor that you have definitely proved that a globulin is responsible for insulin binding, nor that insulin is an antigen. The data you present are indeed suggestive but any more positive claim seems unjustifiable at present.

Sincerely,



Stanley E. Bradley, M.D.
Editor-in-Chief



同行评审 (peer-review)

- ▶ 双盲 (double-blind)、单盲 (single-blind)
- ▶ 审稿人的特征
 - ▶ 可能是大同行、也可能是你的竞争对手
 - ▶ 审稿人是义务的、(带偏见的) 公正、**都很忙**
- ▶ 大部分审稿人的阅读流程:
 - ▶ 5分钟的快速浏览: Title → Abstract → Introduction → Related Work → Conclusion
 - ▶ 做出判断: reject or accept
 - ▶ 验证自己的判断: Proposed Model/Method → Experiment



审稿标准 (AAAI 2018)

- ▶ [Relevance] Is this paper relevant to an AI audience?
- ▶ [Significance] Are the results significant?
- ▶ [Novelty] Are the problems or approaches novel?
- ▶ [Soundness] Is the paper technically sound?
- ▶ [Evaluation] Are claims well-supported by theoretical analysis or experimental results?
- ▶ [Clarity] Is the paper well-organized and clearly written?
- ▶ [OVERALL SCORE]



审稿标准 (EMNLP 2018)

- ▶ **What is this paper about, and what **contributions** does it make?**
Please describe what problem or question this paper addresses, and the main contributions that it makes towards a solution or answer.
- ▶ **What **strengths** does this paper have?**
Please describe the main strengths you see in the paper or the work it describes, regardless of whether you recommend this paper be accepted or not.
- ▶ **What **weaknesses** does this paper have?**
Please describe any weaknesses you see in the paper or the work it describes, regardless of whether you recommend this paper be accepted or not.



审稿标准 (EMNLP 2018为例)

► Overall recommendation

Do you think this paper should be accepted to EMNLP 2018?

In making your overall recommendation, please take into account all of the paper's **strengths and weaknesses**. Please rank short papers relative to other short papers, and long papers relative to other long papers. Acceptable short submissions include: **small, focused contributions; works in progress; negative results and opinion pieces; and interesting application notes**.

- 5 = Exciting: I would fight for this paper to be accepted.
- 4 = Strong: I would like to see it accepted.
- 3 = Borderline: It has some merits but also some serious problems. I'm ambivalent about this one.
- 2 = Mediocre: I would rather not see it in the conference.
- 1 = Poor: I would fight to have it rejected.



审稿标准 (EMNLP 2018为例)

▶ Questions for the Author(s)

Please write any questions you have for the author(s) that you would like answers for in the author response (which you should take into account in your final review).

▶ Missing References

Please list any references that should be included in the bibliography or need to be discussed in more depth.

▶ Presentation Improvements

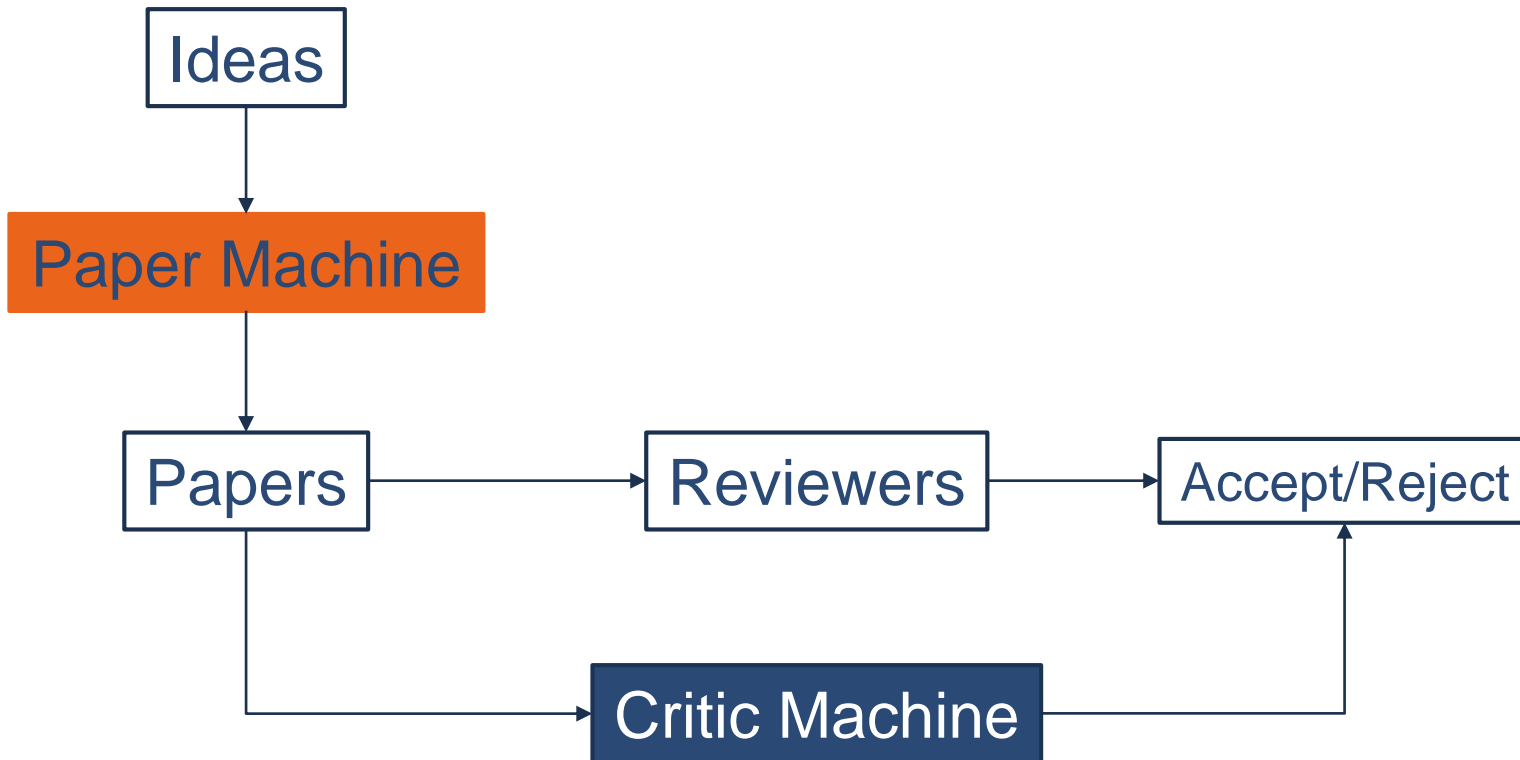
If there is anything in the paper that you found difficult to follow, please suggest how it could be better organized, motivated, or explained.

▶ Typos, Grammar, and Style

Please list any typographical or grammatical errors, as well as any stylistic issues that should be improved.



"Paper Machine" Learning





数据集

A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications

Dongyeop Kang¹ Waleed Ammar² Bhavana Dalvi Mishra² Madeleine van Zuylen²
Sebastian Kohlmeier² Eduard Hovy¹ Roy Schwartz^{2,3}

¹School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

²Allen Institute for Artificial Intelligence, Seattle, WA, USA

³Paul G. Allen Computer Science & Engineering, University of Washington, Seattle, WA, USA
{dongyeok, hovy}@cs.cmu.edu

{waleeda, bhavanad, madeleinev, sebastiank, roys}@allenai.org

Section	#Papers	#Reviews	Asp.	Acc / Rej
NIPS 2013–2017	2,420	9,152	×	2,420 / 0
ICLR 2017	427	1,304	✓	172 / 255
ACL 2017	137	275	✓	88 / 49
CoNLL 2016	22	39	✓	11 / 11
arXiv 2007–2017	11,778	—	—	2,891 / 8,887
<i>total</i>	14,784	10,770		

Asp. Indicates whether the reviews have aspect specific scores (e.g., clarity).

Acc/Rej is the distribution of accepted/rejected papers.



模型

- ▶ We train a binary classifier to estimate the probability of accept vs. reject given a paper,
 - ▶ i.e., $P(\text{accept}=\text{True} \mid \text{paper})$.
- ▶ 模型
 - ▶ Logistic regression,
 - ▶ SVM with linear or RBF kernels,
 - ▶ Random Forest,
 - ▶ Nearest Neighbors,
 - ▶ Decision Tree,
 - ▶ Multilayer Perceptron,
 - ▶ AdaBoost, and
 - ▶ Naive Bayes.

implement all models using sklearn (Pedregosa et al., 2011) with default hyperparameters.



特征

► coarse and lexical features

	Features	Description	Labels
coarse	abstract_contains_X	Whether abstract contains keywords $X \subset \{\text{deep, neural, embedding, outperform, outperform, novel, state_of_the_art}\}$	boolean
	title_length	Length of title	integer
	num_authors	Number of authors	integer
	most_recent_refs_year	Most recent reference year	2001-2017
	num_refs	Number of references (<i>sp</i>)	integer
	num_refmentions	Number of reference mentioned (<i>sp</i>)	integer
	avg_length_refs_mention	Average length of references mentioned (<i>sp</i>)	float
	num_recent_refs	Number of recent references since the paper submitted (<i>sp</i>)	integer
	num_ref_to_X	Number of $X \subset \{\text{figures, tables, sections, equations, theorems}\}$ (<i>sp</i>)	integer
	num_uniq_words	Number of unique words (<i>sp</i>)	integer
	num_sections	Number of sections (<i>sp</i>)	integer
	avg_sentence_length	Average sentence length (<i>sp</i>)	float
	contains_appendix	Whether contains an appendix or not (<i>sp</i>)	boolean
	prop_of_freq_words	Proportion of frequent words (<i>sp</i>)	float
Lexical	BOW	Bag-of-words in abstract	integer
	BOW+TFIDF	TFIDF weighted BOW in abstract	float
	GloVe	Average of GloVe word embeddings in abstract	float
	GloVe+TFIDF	TFIDF weighted average of word embeddings in abstract	float



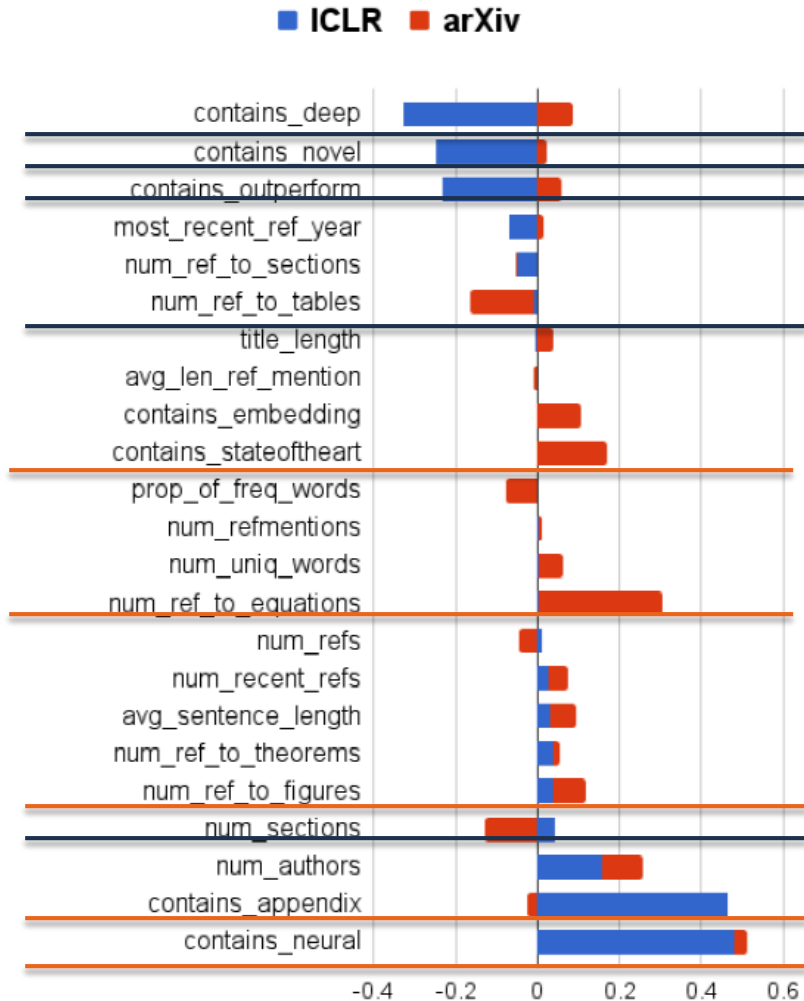
预测结果

- ▶ conduct experiment with the ICLR 2017 and the arXiv sections of the PeerRead dataset.
- ▶ train separate models for each of the arXiv category:
 - ▶ cs.cl, cs.lg, and cs.ai.

	ICLR	cs.cl	cs.lg	cs.ai
Majority	57.6	68.9	67.9	92.1
Ours	65.3	75.7	70.7	92.6
(Δ)	+7.7	+6.8	+2.8	+0.5



Coefficient values for coarse features



▶ a large contribution:

- ▶ adding an appendix,
- ▶ A large number of theorems or equations,
- ▶ the average length of the text preceding a citation,
- ▶ the number of papers cited by this paper that were published in the five years before the submission of this paper,
- ▶ whether the abstract contains a phrase “state of the art” for ICLR or “neural” for arXiv,
- ▶ length of title



分项打分

- ▶ pair-wise correlations between the overall recommendation and various aspect scores in the ACL 2017

Aspect	ρ
Substance	0.59
Clarity	0.42
Appropriateness	0.30
Impact	0.16
Meaningful comparison	0.15
Originality	0.08
Soundness/Correctness	0.01

- ▶ **substance** (which concerns the amount of work rather than its quality)
- ▶ **clarity** (make the paper more easier to read)
- ▶ **soundness/correctness** and **originality** are least correlated with the final recommendation.

Pearson's correlation coefficient



oral vs. poster

- ▶ The average overall recommendation score in reviews recommending an oral presentation is 0.9 higher than in reviews recommending a poster presentation,
- ▶ suggesting that reviewers tend to recommend oral presentation for submissions which are holistically stronger.

Presentation format	Oral	Poster	Δ	stdev
Recommendation	3.83	2.92	0.90	0.89
Substance	3.91	3.29	0.62	0.84
Clarity	4.19	3.72	0.47	0.90
Meaningful comparison	3.60	3.36	0.24	0.82
Impact	3.27	3.09	0.18	0.54
Originality	3.91	3.88	0.02	0.87
Soundness/Correctness	3.93	4.18	-0.25	0.91



ACL 2017 vs. ICLR 2017

Measurement	ACL'17	ICLR'17
Review length (words)	531±323	346±213
Appropriateness	4.9±0.4	2.6±1.3
Meaningful comparison	3.5±0.8	2.9±1.1
Substance	3.6±0.8	3.0±0.9
Originality	3.9±0.9	3.3±1.1
Clarity	3.9±0.9	4.2±1.0
Impact	3.2±0.5	3.4±1.0
Overall recommendation	3.3±0.9	3.3±1.4



总结：论文录用的关键特征

- ▶ 表达清晰，提高读者阅读时的愉悦性
 - ▶ 逻辑性强、容易理解
 - ▶ 合理利用图、表、公式
 - ▶ 排版美观、赏心悦目
 - ▶ 避免低级的语法错误、排版错误、拼写错误
 - ▶ 通过附录展示更详尽的信息，显著提高论文可读性

- ▶ 工作扎实
 - ▶ 有一定的工作量
 - ▶ 实验设计合理、实验详实
 - ▶ State-of-the-art结果

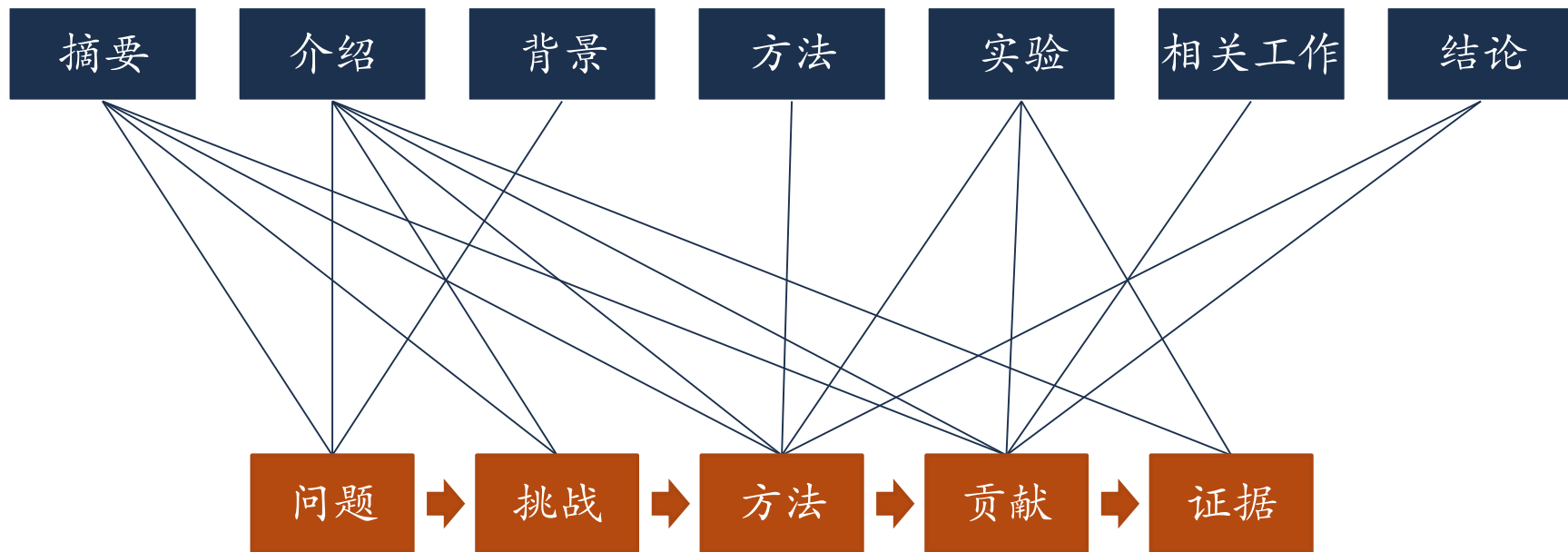


错误的写作方式

- ▶ 内容越**多**越好
- ▶ 题目越**大**越好
- ▶ 方法越**复杂**越好
- ▶ 过度使用模糊词汇，比如“大概”什么的
- ▶ 写作越花哨越好
 - ▶ Overclaimed
- ▶ 糊弄审稿人



科研论文的“八股文”风格





下面的内容拷贝（**略加修改**）自**清华大学刘洋老师**的《学术论文写作技巧》



“摘要”的写法

- ▶ 几句话概括你的工作
- ▶ 误区
 - ▶ 力图把所有细节都说清楚
 - ▶ 用很专业的术语来描述
 - ▶ 出现数学符号

用语要简单，让外行能看懂



例子

Abstract

问题是什么?

Different linguistic perspectives causes many diverse segmentation criteria for Chinese word segmentation (CWS). Most existing methods focus on improve the performance for each single criterion. However, it is interesting to exploit these different criteria and mining their common underlying knowledge. In this paper, we propose adversarial multi-criteria learning for CWS by integrating shared knowledge from multiple heterogeneous segmentation criteria. Experiments on eight corpora with heterogeneous segmentation criteria show that the performance of each corpus obtains a significant improvement, compared to single-criterion learning. Source codes of this paper are available on Github¹.

我们要做什么?

大概怎么做的?

做得还不错!

Xinchi Chen, Zhan Shi, Xipeng Qiu, Xuanjing Huang. Adversarial Multi-Criteria Learning for Chinese Word Segmentation, ACL, 2017. Outstanding Paper Award



“介绍”的写法

- ▶ 比摘要更进一步，用几段话说清你的工作
 - ▶ 充分论证你所做工作的必要性和重要性，要让审稿人认同并迫不及待想往下看。
- ▶ 逻辑性强，论证充分
 - ▶ 起承转合，层层递进



Richard Dawkins 《解析彩虹》

我们都会死，因此都是幸运儿。

绝大多数人永不会死，因为他们从未出生。

那些本有可能取代我的位置但事实上从未见过天日的人，数量多过阿拉伯的沙粒。

那些从未出生的魂灵中，定然有超越济慈的诗人、比牛顿更卓越的科学家。

DNA组合所允许的人类之数，远远超过曾活过的所有人数。

你和我，尽管如此平凡，但仍从这概率低得令人眩晕的命运利齿下逃脱，来到世间。

Xinchi Chen, Zhan Shi, Xipeng Qiu, Xuanjing Huang

Shanghai Key Laboratory of Intelligent Information Processing, Fudan University

School of Computer Science, Fudan University

825 Zhangheng Road, Shanghai, China

{xinchichen13,zshi16,xpqiuxjhuang}@fudan.edu.cn

例子

Abstract

Corpora	Yao	Ming	reaches	the final
CTB		姚明	进入	总决赛
PKU	姚	明	进入	总 决赛

Table 1: Illustration of the different segmentation criteria.

承: 相关工作
 Recently, some efforts have been made to exploit heterogeneous annotation data for Chinese word segmentation or part-of-speech tagging (Jiang et al., 2009; Sun and Wan, 2012; Qiu et al., 2013; Li et al., 2015, 2016). These methods adopted stacking or multi-task architectures and showed that heterogeneous corpora can help each other. However, most of these models adopt the shallow linear classifier with discrete features, which make it difficult to design the shared feature spaces, usually used in a complex model. Fortunately, recent deep neural models provide a convenient way to share information among multiple tasks (Collobert and Weston, 2008; Luong et al., 2015; Chen et al., 2016).

转: 相关工作的不足
 与转机
 In this paper, we propose an adversarial multi-criteria learning for CWS by integrating shared knowledge from multiple segmentation criteria. Specifically, we regard each segmentation criterion as a single task and propose three different shared-private models under the framework of multi-task learning (Caruana, 1997; Ben-David and Schuller, 2003), where a shared layer is used to extract the criteria-invariant features, and a private layer is used to extract the criteria-specific features. Inspired by the success of adversarial strategy on domain adaption (Ajakan et al., 2014; Ganin et al., 2016; Bousmalis et al., 2016), we further utilize adversarial strategy to make sure the shared layer can extract the common underlying and criteria-invariant features, which are suitable for all the criteria. Finally, we exploit the eight segmentation criteria on the five simplified Chi-

1 Introduction

起: 问题是什么?
 Chinese word segmentation (CWS) is a preliminary and important foundation for natural language processing (NLP). Currently, the state-of-the-art methods are based on statistical supervised learning algorithms, and rely on a large-scale annotated corpus whose cost is extremely expensive. Although there have been great achievements in building CWS corpora, they are somewhat incompatible due to different segmentation criteria. As shown in Table 1, given a sentence “姚明进入总决赛 (YaoMing reaches the final)”, the two commonly-used corpora, PKU’s People’s Daily (PKU) (Yu et al., 2001) and Penn Chinese Treebank (CTB) (Fei, 2000), use different segmentation criteria. In a sense, it is a waste of resources if we fail to fully exploit these corpora.

*Corresponding author.
¹<https://github.com/FudanNLP>

nese and three traditional Chinese corpora. Experiments show that our models are effective to improve the performance for CWS. We also observe that traditional Chinese could benefit from incorporating knowledge from simplified Chinese.

The contributions of this paper could be summarized as follows.

- Multi-criteria learning is first introduced for CWS, in which we propose three shared-private models to integrate multiple segmentation criteria.
- An adversarial strategy is used to force the shared layer to learn criteria-invariant features, in which a new objective function is also proposed instead of the original cross-entropy loss.
- We conduct extensive experiments on eight CWS corpora with different segmentation criteria, which is by far the largest number of datasets used simultaneously.

2 General Neural Model for Chinese Word Segmentation

Chinese word segmentation task is usually regarded as a character based sequence labeling problem. Specifically, each character in a sentence is labeled as one of $\mathcal{L} = \{B, M, E, S\}$, indicating the begin, middle, end of a word, or a word with single character. There are lots of prevalent methods to solve sequence labeling problem such as maximum entropy Markov model (MEMM), conditional random fields (CRF), etc. Recurrent neural networks are widely applied to Chinese word segmentation task for their ability to minimize the effort in feature engineering (Zheng et al., 2013; Pei et al., 2014; Chen et al., 2015a,b).

Specifically, given a sequence with n characters $X = \{x_1, \dots, x_n\}$, the aim of CWS task is to figure out the ground truth of labels $Y^* = \{y_1^*, \dots, y_n^*\}$:

$$Y^* = \arg \max_{Y \in \mathcal{L}^n} p(Y|X), \quad (1)$$

where $\mathcal{L} = \{B, M, E, S\}$.

The general architecture of neural CWS could be characterized by three components: (1) a character embedding layer; (2) feature layers consisting of several classical neural networks and (3) a tag inference layer. The role of feature layers is to extract features, which could be either convolution neural network or recurrent neural network. In this

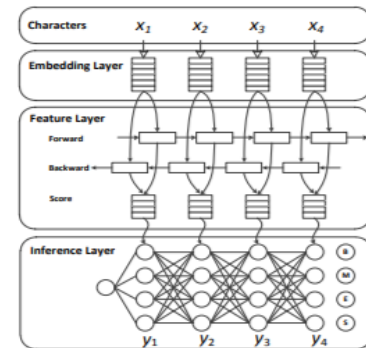


Figure 1: General neural architecture for Chinese word segmentation.

paper, we adopt the bi-directional long short-term memory neural networks followed by CRF as the tag inference layer. Figure 1 illustrates the general architecture of CWS.

2.1 Embedding layer

In neural models, the first step usually is to map discrete language symbols to distributed embedding vectors. Formally, we lookup embedding vector from embedding matrix for each character x_i as $e_{x_i} \in \mathbb{R}^{d_e}$, where d_e is a hyper-parameter indicating the size of character embedding.

2.2 Feature layers

We adopt bi-directional long short-term memory (BiLSTM) as feature layers. While there are numerous LSTM variants, here we use the LSTM architecture used by (Jozefowicz et al., 2015), which is similar to the architecture of (Graves, 2013) but without peep-hole connections.

LSTM LSTM introduces gate mechanism and memory cell to maintain long dependency information and avoid gradient vanishing. Formally, LSTM, with input gate \mathbf{i} , output gate \mathbf{o} , forget gate \mathbf{f} and memory cell \mathbf{c} , could be expressed as:

$$\begin{bmatrix} \mathbf{i} \\ \mathbf{o} \\ \mathbf{f} \\ \mathbf{c} \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \phi \end{bmatrix} \left(\mathbf{W}_g^\top \begin{bmatrix} \mathbf{e}_{x_i} \\ \mathbf{h}_{i-1} \end{bmatrix} + \mathbf{b}_g \right), \quad (2)$$

$$\mathbf{c}_i = \mathbf{c}_{i-1} \odot \mathbf{f}_i + \mathbf{c}_i \odot \mathbf{i}_i, \quad (3)$$

$$\mathbf{h}_i = \mathbf{o}_i \odot \phi(\mathbf{c}_i), \quad (4)$$

where $\mathbf{W}_g \in \mathbb{R}^{(d_e+d_h) \times 4d_h}$ and $\mathbf{b}_g \in \mathbb{R}^{4d_h}$ are trainable parameters. d_h is a hyper-parameter, in-

背景方法介绍



“方法” 的写法

- ▶ 不要一上来就描述你的工作，可以先介绍背景知识
 - ▶ （往往就是baseline）
 - ▶ 有利于降低初学者或其他领域学者的理解难度
 - ▶ 有利于对introduction中的论证做更详细的解释
 - ▶ 有利于对比baseline和你的方法

- ▶ 合理使用图、表、公式
- ▶ 使用一个running example来阐释你的方法
- ▶ 逻辑清楚、可重复

3 Multi-Criteria Learning for Chinese Word Segmentation

Although neural models are widely used on CWS, most of them cannot deal with incompatible criteria with heterogenous segmentation criteria simultaneously.

Inspired by the success of multi-task learning (Caruana, 1997; Ben-David and Schuller, 2003; Liu et al., 2016a,b), we regard the heterogenous criteria as multiple “related” tasks, which could improve the performance of each other simultaneously with shared information.

Formally, assume that there are M corpora with heterogeneous segmentation criteria. We refer \mathcal{D}_m as corpus m with N_m samples:

$$\mathcal{D}_m = \{(X_i^{(m)}, Y_i^{(m)})\}_{i=1}^{N_m}, \quad (11)$$

where X_i^m and Y_i^m denote the i -th sentence and the corresponding label in corpus m .

To exploit the shared information between these different criteria, we propose three sharing models for CWS task as shown in Figure 2. The feature layers of these three models consist of a private (criterion-specific) layer and a shared (criterion-invariant) layer. The difference between three models is the information flow between the task layer and the shared layer. Besides, all of these three models also share the embedding layer.

简要地重复问题

解决思路

必要的形式化定义

公式的正确写法:

1. 公式都要加编号
2. 新出现的符号都要进行说明
3. 避免重复使用同一符号
4. 遵守约定俗成的表示方式
5. 合理地使用 $,/.$
6. **where** 不缩进
7. 合理使用上下标

具体模型

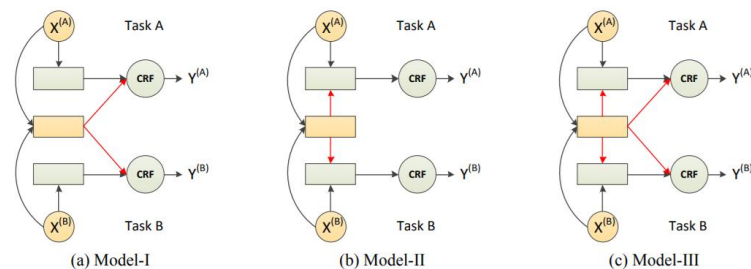


Figure 2: Three shared-private models for multi-criteria learning. The yellow blocks are the shared Bi-LSTM layer, while the gray block are the private Bi-LSTM layer. The yellow circles denote the shared embedding layer. The red information flow indicates the difference between three models.

简要的模型可视化

也可以是和baseline模型对比



“实验”的写法

- ▶ 关键
 - ▶ 给论文贡献提供证据
 - ▶ 扎实 (solid) 可靠 (reliable)

- ▶ 实验设计非常重要
 - ▶ 想读者之所想，将读者关心的问题事先想好
 - ▶ 公认的标准数据和state-of-the-art系统
 - ▶ 实验层层递进，有逻辑性
 - ▶ 主实验（测试集）：证明显著超过baseline

- ▶ 不辞辛劳，做到极致



数据集和实验设置

必要时可以通过验证集上的辅助实验来说明为什么设置这些参数

例子

6 Experiments

6.1 Datasets

To evaluate our proposed architecture, we experiment on eight prevalent CWS datasets from SIGHAN2005 (Emerson, 2005) and SIGHAN2008 (Jin and Chen, 2008). Table 2 gives the details of the eight datasets. Among these datasets, AS, CITYU and CKIP are traditional Chinese, while the remains, MSRA, PKU, CTB, NCC and SXU, are simplified Chinese. We use 10% data of shuffled train set as development set for all datasets.

Datasets			Words	Chars	Word Types	Char Types	Sents	OOV Rate
Sighan05	MSRA	Train	2.4M	4.1M	88.1K	5.2K	86.9K	-
		Test	0.1M	0.2M	12.9K	2.8K	4.0K	2.60%
	AS	Train	5.4M	8.4M	141.3K	6.1K	709.0K	-
		Test	0.1M	0.2M	18.8K	3.7K	14.4K	4.30%
Sighan08	PKU	Train	1.1M	1.8M	55.2K	4.7K	47.3K	-
		Test	0.2M	0.3M	17.6K	3.4K	6.4K	-
	CTB	Train	0.6M	1.1M	42.2K	4.2K	23.4K	-
		Test	0.1M	0.1M	9.8K	2.6K	2.1K	5.55%
	CKIP	Train	0.7M	1.1M	48.1K	4.7K	94.2K	-
		Test	0.1M	0.1M	15.3K	3.5K	10.9K	7.41%
	CITYU	Train	1.1M	1.8M	43.6K	4.4K	36.2K	-
		Test	0.2M	0.3M	17.8K	3.4K	6.7K	8.23%
	NCC	Train	0.5M	0.8M	45.2K	5.0K	18.9K	-
		Test	0.1M	0.2M	17.5K	3.6K	3.6K	4.74%
	SXU	Train	0.5M	0.9M	32.5K	4.2K	17.1K	-
		Test	0.1M	0.2M	12.4K	2.8K	3.7K	5.12%

Table 2: Details of the eight datasets.

6.2 Experimental Configurations

For hyper-parameter configurations, we set both the character embedding size d_e and the dimensionality of LSTM hidden states d_h to 100. The initial learning rate α is set to 0.01. The loss weight coefficient λ is set to 0.05. Since the scale of each dataset varies, we use different training batch sizes for datasets. Specifically, we set batch sizes of AS and MSR datasets as 512 and 256 respectively, and 128 for remains. We employ dropout strategy on embedding layer, keeping 80% inputs (20% dropout rate).

For initialization, we randomize all parameters following uniform distribution at $(-0.05, 0.05)$. We simply map traditional Chinese characters to simplified Chinese, and optimize on the same character embedding matrix across datasets, which is pre-trained on Chinese Wikipedia corpus, using word2vec toolkit (Mikolov et al., 2013). Following previous work (Chen et al., 2015b; Pei et al., 2014), all experiments including baseline results are using pre-trained character embedding with bi-gram feature.

主实验

6.3 Overall Results

Table 3 shows the experiment results of the proposed models on test sets of eight CWS datasets, which has three blocks.

(1) In the first block, we can see that the performance is boosted by using Bi-LSTM, and the performance of Bi-LSTM cannot be improved by merely increasing the depth of networks. In addition, although the F value of LSTM model in (Chen et al., 2015b) is 97.4%, they additionally incorporate an external idiom dictionary.

(2) In the second block, our proposed three models based on multi-criteria learning boost performance. Model-I gains 0.75% improvement on averaging F-measure score compared with Bi-LSTM result (94.14%). Only the performance on MSRA drops slightly. Compared to the baseline results (Bi-LSTM and stacked Bi-LSTM), the proposed models boost the performance with the help of exploiting information across these heterogeneous segmentation criteria. Although various criteria have different segmentation granularities, there are still some underlying information shared. For instance, MSRA and CTB treat family name and last name as one token “宁泽涛 (NingZeTao)”, whereas some other datasets, like PKU, regard them as two tokens, “宁 (Ning)” and “泽涛 (Ze-Tao)”. The partial boundaries (before “宁 (Ning)” or after “涛 (Tao)”) can be shared.

(3) In the third block, we introduce adversarial training. By introducing adversarial training, the performances are further boosted, and Model-I is slightly better than Model-II and Model-III. The adversarial training tries to make shared layer keep criteria-invariant features. For instance, as shown in Table 3, when we use shared information, the performance on MSRA drops (worse than baseline result). The reason may be that the shared parameters bias to other segmentation criteria and introduce noisy features into shared parameters. When we additionally incorporate the adversarial training

详尽的
实验分
析

Models		MSRA	AS	PKU	CTB	CKIP	CITYU	NCC	SXU	Avg.
LSTM	P	95.13	93.66	93.96	95.36	91.85	94.01	91.45	95.02	93.81
	R	95.55	94.71	92.65	85.52	93.34	94.00	92.22	95.05	92.88
	F	95.34	94.18	93.30	95.44	92.59	94.00	91.83	95.04	93.97
	OOV	63.60	67.83	67.81	67.81	67.81	67.48	56.28	69.46	67.00
Bi-LSTM	P	95.70	94.20	95.30	95.30	95.06	94.00	91.86	95.11	93.95
	R	95.99	70.07	66.09	76.47	72.12	65.15	92.47	95.23	94.33
	F	95.84	70.07	66.09	76.47	72.12	65.15	92.17	95.17	94.14
	OOV	66.28	70.07	66.09	76.47	72.12	65.79	59.11	71.27	68.40
Stacked Bi-LSTM	P	95.69	93.89	94.10	95.20	92.40	94.13	91.81	94.99	94.03
	R	95.81	94.54	92.66	95.40	93.39	93.99	92.62	95.37	94.22
	F	95.75	94.22	93.37	95.30	92.89	94.06	92.21	95.18	94.12
	OOV	65.55	71.50	67.92	75.44	70.50	66.35	57.39	69.69	68.04
Multi-Criteria Learning										
Model-I	P	95.67	94.44	94.93	95.95	93.99	95.10	92.54	96.07	94.84
	R	95.82	95.09	93.73	96.00	94.52	95.60	92.69	96.08	94.94
	F	95.74	94.76	94.33	95.97	94.26	95.35	92.61	96.07	94.89
	OOV	69.89	74.13	72.96	81.12	77.58	80.00	64.14	77.05	74.61
Model-II	P	95.74	94.52	94.65	96.09	93.80	95.37	92.26	96.17	94.79
	R	95.74	74.87	72.28	79.94	76.67	81.05	92.84	95.95	94.94
	F	95.74	74.87	72.28	79.94	76.67	81.05	92.55	96.06	94.86
	OOV	69.67	74.87	72.28	79.94	76.67	81.05	61.51	77.96	74.24
Model-III	P	95.76	93.99	94.95	95.85	93.50	95.56	92.17	96.10	94.74
	R	95.89	95.07	93.48	96.11	94.58	95.62	92.96	96.13	94.98
	F	95.82	94.53	94.21	95.98	94.04	95.59	92.57	96.12	94.86
	OOV	70.72	72.59	73.12	81.21	76.56	82.14	60.83	77.56	74.34
Adversarial Multi-Criteria Learning										
Model-I+ADV	P	95.95	94.17	94.86	96.02	93.82	95.39	92.46	96.07	94.84
	R	96.14	95.11	93.78	96.33	94.70	95.70	93.19	96.01	95.12
	F	96.04	94.64	94.32	96.18	94.26	95.55	92.83	96.04	94.98
	OOV	71.60	73.50	72.67	82.48	77.59	81.40	63.31	77.10	74.96
Model-II+ADV	P	96.02	94.52	94.65	96.09	93.80	95.37	92.42	95.85	94.84
	R	95.86	73.57	73.13	82.13	77.71	81.05	93.20	96.07	94.99
	F	95.94	73.57	73.13	82.13	77.71	81.05	92.81	95.96	94.92
	OOV	72.76	73.57	73.13	82.13	77.71	81.05	62.16	76.88	75.16
Model-III+ADV	P	95.92	94.25	94.68	95.86	93.67	95.24	92.47	96.24	94.79
	R	95.83	95.11	93.82	96.10	94.48	95.60	92.73	96.04	94.96
	F	95.87	94.68	94.25	95.98	94.07	95.42	92.60	96.14	94.88
	OOV	70.86	72.89	72.20	81.65	76.13	80.71	63.22	77.88	74.44

Table 3: Results of proposed models on test sets of eight CWS datasets. There are three blocks. The first block consists of two baseline models: Bi-LSTM and stacked Bi-LSTM. The second block consists of our proposed three models without adversarial training. The third block consists of our proposed three models with adversarial training. Here, P, R, F, OOV indicate the precision, recall, F value and OOV recall rate respectively. The maximum F values in each block are highlighted for each dataset.

在caption中描述必要的信息

6.4 Speed

模型速度和错误分析

To further explore the convergence speed, we plot the results on development sets through epochs. Figure 4 shows the learning curve of Model-I without incorporating adversarial strategy. As shown in Figure 4, the proposed model makes progress gradually on all datasets. After about 1000 epochs, the performance becomes stable and convergent.

We also test the decoding speed, and our models process 441.38 sentences per second averagely. As the proposed models and the baseline models (Bi-LSTM and stacked Bi-LSTM) are nearly in the same complexity, all models are nearly the same efficient. However, the time consumption of training process varies from model to model. For the models without adversarial training, it costs about 10 hours for training (the same for stacked Bi-LSTM to train eight datasets), whereas it takes about 16 hours for the models with adversarial training. All the experiments are conducted on the hardware with Intel(R) Xeon(R) CPU E5-2643 v3 @ 3.40GHz and NVIDIA GeForce GTX TITAN X.

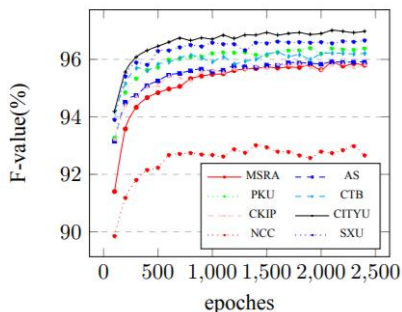


Figure 4: Convergence speed of Model-I without adversarial training on development sets of eight datasets.

6.5 Error Analysis

We further investigate the benefits of the proposed models by comparing the error distributions between the single-criterion learning (baseline model Bi-LSTM) and multi-criteria learning (Model-I and Model-I with adversarial training) as shown in Figure 5. According to the results, we could observe that a large proportion of points lie above diagonal lines in Figure 5a and Figure 5b, which implies that performance benefit from integrating knowledge and complementary information from other corpora. As shown in Table 3, on the test set of CITYU, the performance of Model-I and its adversarial version (Model-I+ADV) boost from 92.17% to 95.59% and 95.42% respectively.

In addition, we observe that adversarial strategy is effective to prevent criterion specific features from creeping into shared space. For instance, the segmentation granularity of personal name is often different according to heterogenous criteria. With the help of adversarial strategy, our models could correct a large proportion of mistakes on personal name. Table 4 lists the examples from 2333-th and 89-th sentences in test sets of PKU and MSRA datasets respectively.

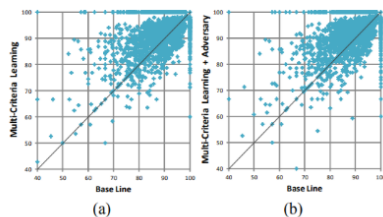


Figure 5: F-measure scores on test set of CITYU dataset. Each point denotes a sentence, with the (x, y) values of each point denoting the F-measure scores of the two models, respectively. (a) is comparison between Bi-LSTM and Model-I. (b) is comparison between Bi-LSTM and Model-I with adversarial training.

Models	PKU-2333		MSRA-89
Golds	Roh 卢	Moo-hyun 武铉	Mu Ling Ying 穆玲英
Base Line	卢武铉		穆 玲英
Model-I	卢武铉		穆 玲英
Modell-I+ADV	卢	武铉	穆玲英

Table 4: Segmentation cases of personal names.

7 Knowledge Transfer 迁移能力

We also conduct experiments of whether the shared layers can be transferred to the other related tasks or domains. In this section, we investigate the ability of knowledge transfer on two experiments: (1) simplified Chinese to traditional Chinese and (2) formal texts to informal texts.

7.1 Simplified Chinese to Traditional Chinese

Traditional Chinese and simplified Chinese are two similar languages with slightly difference on character forms (e.g. multiple traditional characters might map to one simplified character). We investigate that if datasets in traditional Chinese and simplified Chinese could help each other. Table 5 gives the results of Model-I on 3 traditional Chinese datasets under the help of 5 simplified Chinese datasets. Specifically, we firstly train the model on simplified Chinese datasets, then we train traditional Chinese datasets independently with shared parameters fixed.

As we can see, the average performance is boosted by 0.41% on F-measure score (from 93.78% to 94.19%), which indicates that shared features learned from simplified Chinese segmentation criteria can help to improve performance on traditional Chinese. Like MSRA, as AS dataset is relatively large (train set of 5.4M tokens), the features learned by shared parameters might bias to other datasets and thus hurt performance on such large dataset AS.

7.2 Formal Texts to Informal Texts

7.2.1 Dataset

We use the NLPCC 2016 dataset² (Qiu et al., 2016) to evaluate our model on micro-blog texts. The NLPCC 2016 data are provided by the shared task in the 5th CCF Conference on Natural Language Processing & Chinese Computing (NLPCC 2016): Chinese Word Segmentation and POS Tagging for micro-blog Text. Unlike the popular used news-wire dataset, the NLPCC 2016 dataset is collected from Sina Weibo³, which consists of the informal texts from micro-blog with the various topics, such as finance, sports, entertainment, and so on. The information of the dataset is shown in Table 6.

7.2.2 Results

Formal documents (like the eight datasets in Table 2) and micro-blog texts are dissimilar in many aspects. Thus, we further investigate that if the formal texts could help to improve the performance of micro-blog texts. Table 7 gives the results of Model-I on the NLPCC 2016 dataset under the help of the eight datasets in Table 2. Specifically, we firstly train the model on the eight datasets, then we train on the NLPCC 2016 dataset alone with shared parameters fixed. The baseline model is Bi-LSTM which is trained on the NLPCC 2016 dataset alone.

As we can see, the performance is boosted by 0.30% on F-measure score (from 93.94% to 94.24%), and we could also observe that the OOV recall rate is boosted by 3.97%. It shows that the shared features learned from formal texts can help to improve the performance on of micro-blog texts.

Models	AS	CKIP	CITYU	Avg.
Baseline(Bi-LSTM)	94.20	93.06	94.07	93.78
Model-I*	94.12	93.24	95.20	94.19

Table 5: Performance on 3 traditional Chinese datasets. Model-I* means that the shared parameters are trained on 5 simplified Chinese datasets and are fixed for traditional Chinese datasets. Here, we conduct Model-I without incorporating adversarial training strategy.

Dataset	Words	Chars	Word Types	Char Types	Sents	OOV Rate
Train	421,166	688,743	43,331	4,502	20,135	-
Dev	43,697	73,246	11,187	2,879	2,052	6.82%
Test	187,877	315,865	27,804	3,911	8,592	6.98%

Table 6: Statistical information of NLPCC 2016 dataset.

Models	P	R	F	OOV
Baseline(Bi-LSTM)	93.56	94.33	93.94	70.75
Model-I*	93.65	94.83	94.24	74.72

Table 7: Performances on the test set of NLPCC 2016 dataset. Model-I* means that the shared parameters are trained on 8 Chinese datasets (Table 2) and are fixed for NLPCC dataset. Here, we conduct Model-I without incorporating adversarial training strategy.



“相关工作”的写法

▶ 错误的写法

- ▶ 没有引用重要论文（可以直接作为rejection的理由）
- ▶ 简单的罗列和堆砌，缺乏深刻到位的评论
- ▶ “贬低”他人工作

▶ 正确的写法

- ▶ 向审稿人显示你对本领域具有全面深刻的把握
- ▶ 通过与前人工作的对比凸显你的工作的创新性
- ▶ 为读者梳理领域的发展脉络，获得全局的认识



例子

改进的工作

不足之处

和直接竞争模型的区别

我们工作的共享和优点

Li et al. (2015) proposed a coupled sequence labeling model which could directly learn and infer two heterogeneous annotations. Chao et al. (2015) also utilize multiple corpora using coupled sequence labeling model. These methods adopt the shallow classifiers, therefore suffering from the problem of defining shared features.

Our proposed models use deep neural networks, which can easily share information with hidden shared layers. Chen et al. (2016) also adopted neural network models for exploiting heterogeneous annotations based on neural multi-view model, which can be regarded as a simplified version of our proposed models by removing private hidden layers.

Unlike the above models, we design three sharing-private architectures and keep shared layer to extract criterion-invariance features by introducing adversarial training. Moreover, we fully exploit eight corpora with heterogeneous segmentation criteria to model the underlying shared information.

8 Related Works

There are many works on exploiting heterogeneous annotation data to improve various NLP tasks. Jiang et al. (2009) proposed a stacking-based model which could train a model for one specific desired annotation criterion by utilizing knowledge from corpora with other heterogeneous annotations. Sun and Wan (2012) proposed a structure-based stacking model to reduce the approximation error, which makes use of structured features such as sub-words. These models are unidirectional aid and also suffer from error propagation problem.

Qiu et al. (2013) used multi-tasks learning framework to improve the performance of POS tag-

承认开创性的工作

不足之处



“标题” 的写法

- ▶ 用一句话概括你所做的工作
 - ▶ 问题+方法+贡献
 - ▶ 可以适当别出心裁

例子: Best Papers of ACL 2018

Best Long Papers

Finding syntax in human encephalography with beam search. John Hale, Chris Dyer, Adhiguna Kuncoro and Jonathan Brennan.

Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information. Sudha Rao and Hal Daumé III.

Let's do it "again": A First Computational Approach to Detecting Adverbial Presupposition Triggers. Andre Cianflone,* Yulan Feng,* Jad Kabbara* and Jackie Chi Kit Cheung. (* equal contribution)

Best Short Papers

Know What You Don't Know: Unanswerable Questions for SQuAD. Pranav Rajpurkar, Robin Jia and Percy Liang

'Lighter' Can Still Be Dark: Modeling Comparative Color Descriptions. Olivia Winn and Smaranda Muresan



美观的排版：读起来赏心悦目、心情舒畅

▶ 需要掌握的工具

▶ 主文档：Latex/XeLatex

▶ <https://www.overleaf.com/>

▶ <https://en.wikibooks.org>

▶ 可视化：Tikz/Gnuplot/MetaPost

▶ <http://www.texample.net/tikz/examples/>

▶ <https://www.overleaf.com/gallery/tagged/tikz>

▶ 协调配色

▶ <https://color.adobe.com/zh/explore>

▶ 参考文献：Bibtex

▶ <https://dblp.org/>

▶ 减少低级语法/拼写错误

▶ 神器：<https://www.grammarly.com/>

▶ 好好利用google!

□ 不确定的写法可以看看别人怎么写的



总结

- ▶ 科研论文是为了分享自己的工作，前提是有“好的工作”
- ▶ 以读者为中心，充分考虑读者的接受能力，提高阅读乐趣
- ▶ 不论在篇章层面还是段落层面，都要具有逻辑性
 - ▶ 起承转合，层层递进
 - ▶ 英语不是关键，逻辑性更重要
- ▶ 实验详尽、扎实可靠
- ▶ 排版美观



Last but not least



合理对待论文被拒

▶ 客观对待评审意见，完善论文！

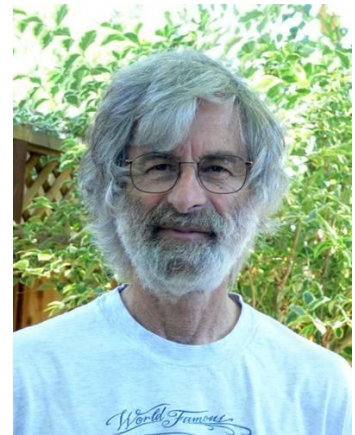
The Part-Time Parliament Paxos算法

ACM Transactions on Computer Systems 16, 2 (May 1998), 133-169. Also appeared as SRC Research Report 49. This paper was first submitted in 1990, setting a personal record for publication delay that has since been broken by [60].

Buridan's Principle

Foundations of Physics 42, 8 (August 2012) 1056-1066. in December of 1984 I wrote this paper.

<http://lamport.azurewebsites.net/pubs/pubs.html>



Leslie Lamport
2013 Turing Award

衣带渐宽终不悔 为伊消得人憔悴



谢 谢