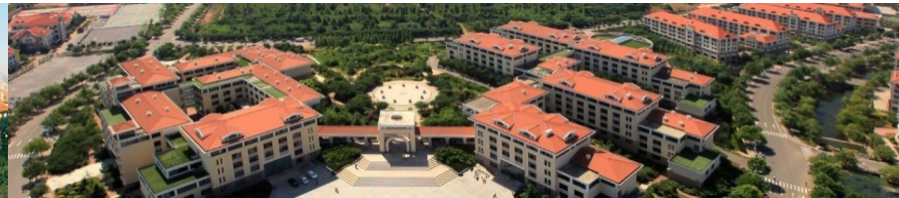
学术论文写作

绪论 写作方法和技巧

2020 / 09



课程主要内容

第一部分：科技论文撰写预备

成为一流科研人员的规律？如何选择研究领域？如何收集相关材料并阅读？如何确定科研方向与选题？

第二部分：科技论文谋划、构成与规范表达

如何谋划和开始一篇科技论文？科技论文构成与规范表达？科技论文插图与表格规范？科技论文公式的规范？如何完成硕士（博士）学位论文？

课程主要内容

第三部分：学术规范指南

什么叫编、著与编著？科技术论文引文规范是什么？

第四部分：科技论文语言规范

科技论文汉语语言特点、各类语病、科技论文句式选择方法、标点符号的规范使用； 科技论文英语规范表达原则；一些标准的写作技巧，定冠词使用; 论文投稿信与修改

课程学习方式

线上交流 + 课堂学习 + 课后作业

课程微信群

20秋学术论文写作

课程大作业

用英文写一篇短文，包括如下几部分：选题的价值和意义，当前研究是如何做的，当前方法的主要问题，你计划如何解决这些问题，你解决方案的主要亮点。

具体要求：

1.字数1500以上，不可抄袭；

2.筛选出学术价值较高的文献（发表于知名会议或者期刊）进行引用，不少于10篇，参考文献格式参照IEEE格式；

3.在Overleaf上写作：https://www.overleaf.com

4.作业最终提交到网盘，到时候会通知

成绩考核

成绩构成（百分制）：

课程大作业 85% + 课堂作业10% + 出勤5%

+ 奖励成绩 - 惩罚成绩

奖励（主动组织主题讨论、主动分享课程相关学习等）

惩罚（上课玩手机、交头结耳、传播负能量等）

投票(匿名) 最多可选1项

设置

参加《学术论文写作》课程的目的?

A 对学术论文写作非常感兴趣

B 尚不明确，也许会感兴趣

C 对学术写作不感兴趣，就是来混混学分

D 反感从事学术工作

提交

Question & Answer

本讲主要内容来自清华大学刘知远

清华大学计算机系自然语言处理实验室，副教授
电邮：liuzy [at] tsinghua.edu.cn
地址：北京市海淀区清华大学FIT大楼4-506，100084
研究兴趣：知识图谱与语义计算，社会计算与计算社会科学

教育工作经历

- 2017年12月 - 至今. 清华大学计算机系，教研系列准聘副教授.
- 2016年8月 - 2017年12月. 清华大学计算机系，教研系列助理教授.
- 2013年12月 - 2016年8月. 清华大学计算机系，助理研究员.
- 2011年8月 - 2013年12月. 清华大学计算机系，博士后.
- 2006年8月 - 2011年7月. 清华大学计算机系，博士.
- 2002年8月 - 2006年7月. 清华大学计算机系，本科.

本讲主要内容来自清华大学刘知远

如何评价刘知远?

刘知远，自然语言处理、机器学习、深度学习（Deep Learning） 话题的优秀回答者

别再邀请我了。。。我就是我，不一样的烟火。。。

国内自然语言处理（NLP）领域，刘知远能排第几?

刘知远

自然语言处理、机器学习、深度学习（Deep Learning） 话题的优秀回答者

205 人赞同了该回答

谢邀。竟然还有这问题，太尴尬，还是让我来终结这个问题：如果是算体重的话，排名还是颇可观的。

本讲主要内容来自清华大学刘知远

如何评价刘知远?

刘知远，自然语言处理、机器学习、深度学习（Deep Learning） 话题的优秀回答者

别再邀请我了。。。我就是我，不一样的烟火。。。

zibuyu9

3月11日 09:15 来自 微博 weibo.com 已编辑

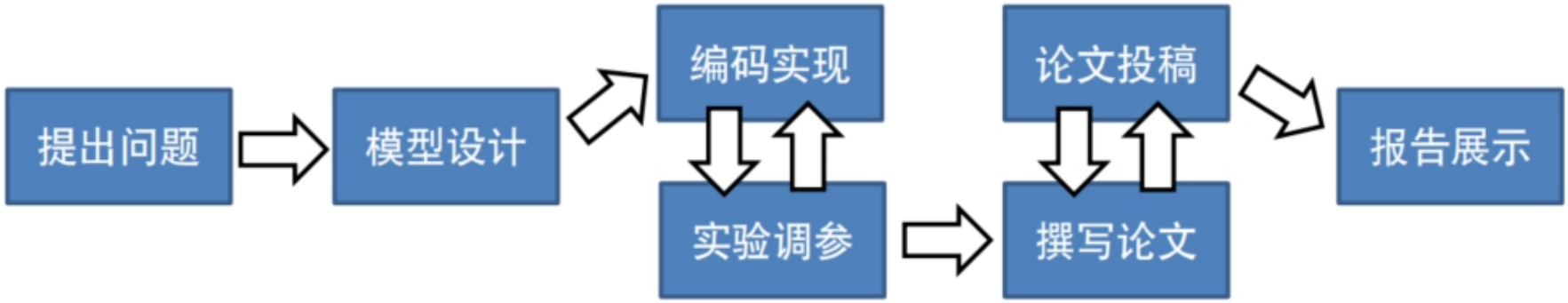
置顶 上周赶完ACL2019用周末时间写了篇儿"如何写好NLP论文"的短文：

如何写一篇合格的NLP论文 不知不觉，这几年已经写了好几篇帮助初学者入门的文章，今天把链接都汇总到个人主页上了，欢迎大家访问。希望帮助更多同学在NLP的科研之路走得更稳更远： 网页链接

学术研究是一项系统工程

计算机学科创新成果可以是新的算法、任务、应用、数据、发现等，务求一个 “新” 字，其影响力则取决于它对该领域发展的推动作用

学术研究是一项系统工程，包括多个环节，共同完成对 “创新” 的追求：**问题务求挑战，模型务求创新，实现务求准确，实验务求深入**



学术研究是一项系统工程

在这个系统工程中，论文的作用则是，向学术界同行清晰准确地描述成果的创新点、技术思路、算法细节和验证结果

明白这一点，才能正确的对待论文写作：

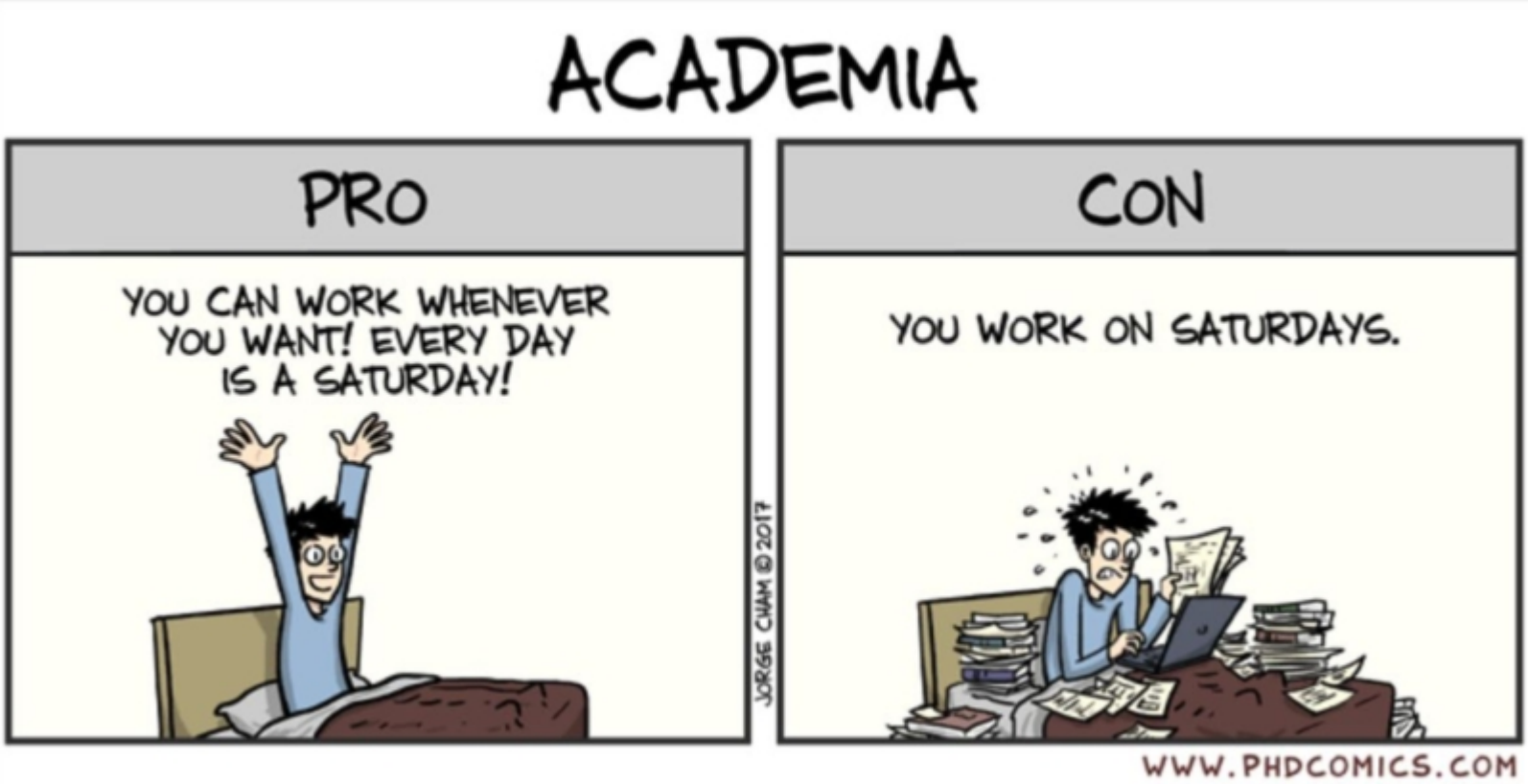
一项乏善可陈的工作，很难通过写作变得众星捧月；

一项充满创新的成果，却有可能因为糟糕的写作而无法向审稿人准确传递重要价值所在，延误成果发表

学术研究需要天时地利人和

成功的研究 =

重要问题 + 新颖的方法 + 努力、积累、坚持



学术研究不同时期有不同的追求

第一层： 锻炼解决开放问题的能力

第二层： 成为相关领域的知名专家

第三层： 做出引领领域方向的工作

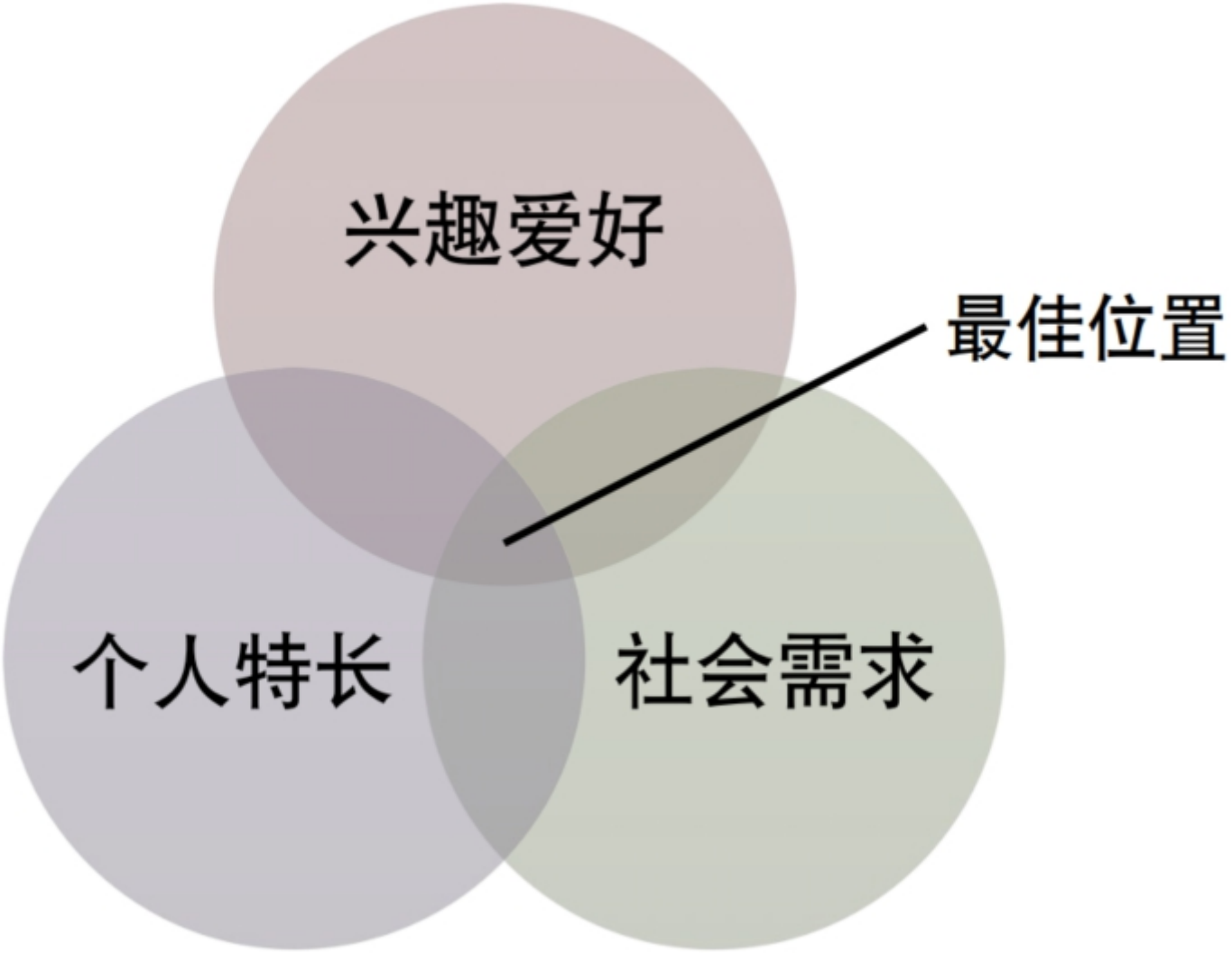
学术研究不同时期有不同的追求

当在发表第一篇论文时，你可能会想如何去解决这个问题，但是我觉得更重要的是要**锻炼问题的研究能力**，因为你要解决的问题尽量要是前人没有解决的

第二层，当博士读到2-3年时候，希望能有两三份重要的工作能使得你成为这个**领域的知名专家**

第三层，你要比绝大多数人**想问题想的更早、眼光更加长远**，这时候你就成为了引领这个领域的专家。不论读博还是读研，希望大家在这三个层次进行进步

研究方向的选择



研究建议1

正视个体差异

扬长避短：不同研究环节侧重不同方面

循序渐进：

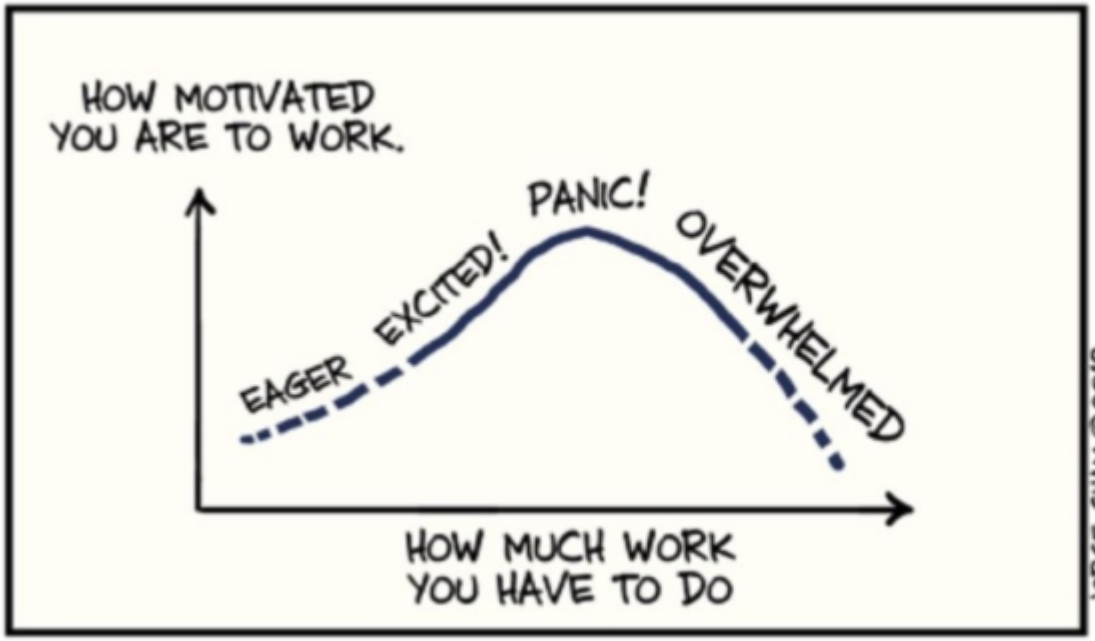
- 从事第一项研究时，要主要负责模型设计与实现，导师主要负责选题、技术路线和论文撰写
- 成功完成首项任务后，则可以开始在选题等方面承担更多责任，从而得到更全面的锻炼

每个人有自己的长处短处，比如你编程好，但英语不好、写作不好。所以要扬长避短，尽量做好自己擅长的，并综合自己的资源

研究建议2

迅速进入研究状态

- 在学习入门知识的同时，迅速从具体研究任务入手，开始研究历练
- 在实践中学习，学以致用，实现对领域的全景了解



研究建议3

把科研列为高优先级

- 历史表明，成绩与重视程度成正比
- 正式加入前，慎重决定，**一旦决定全力以赴**

研究建议3

把科研列为高优先级

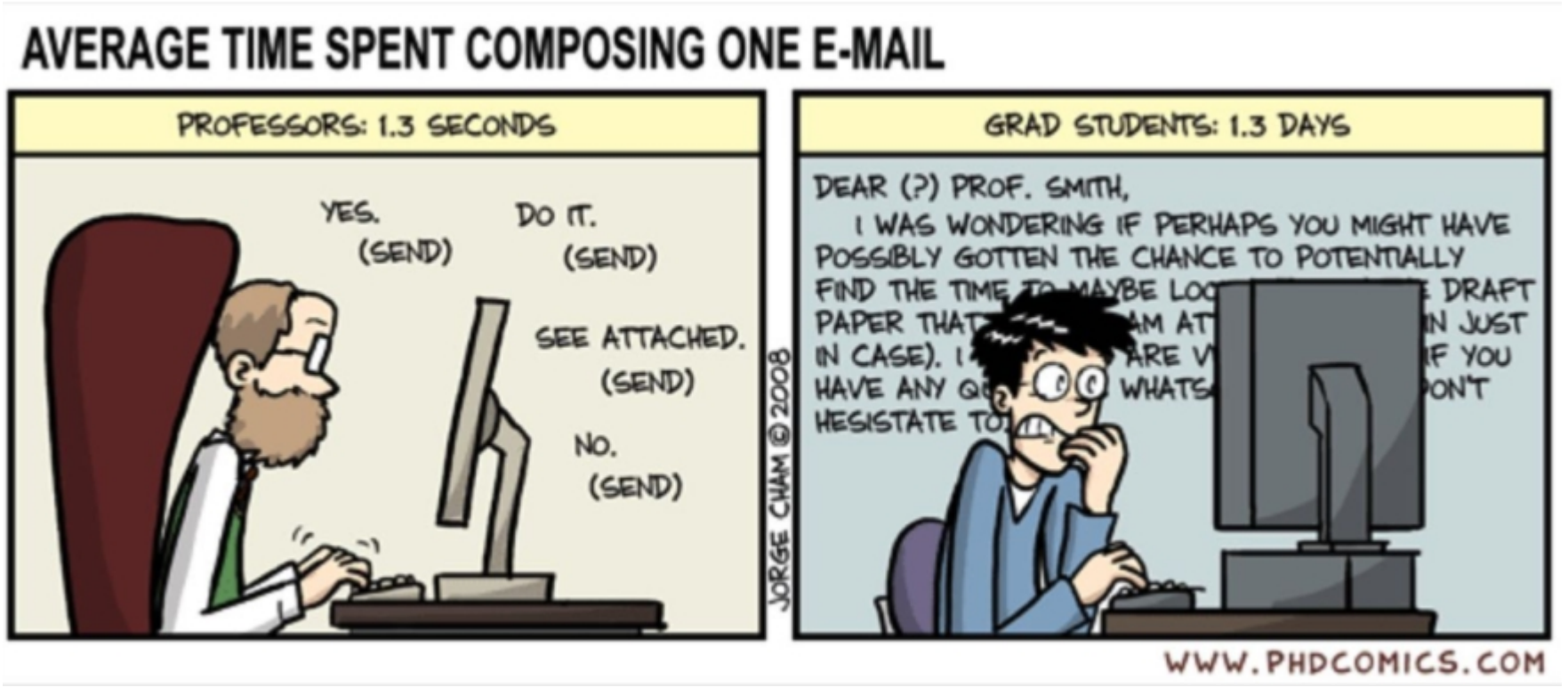
- 历史表明，成绩与重视程度成正比
- 正式加入前，慎重决定，**一旦决定全力以赴**

无论怎么样，如果你想在科研中有所成就，**一定要将科研列为最高优先级**。这种最高优先级并不是每天都在实验室待着，而是一种状态，比如你吃饭时候、洗澡时候、上床睡觉时候都应在想有什么问题没有解决，自己如何进行解决，进行思考

研究建议4

坚持积极主动的态度

- 积极与导师学长交流，充分利用LAB资源，**一切以完成高水平研究为目标**

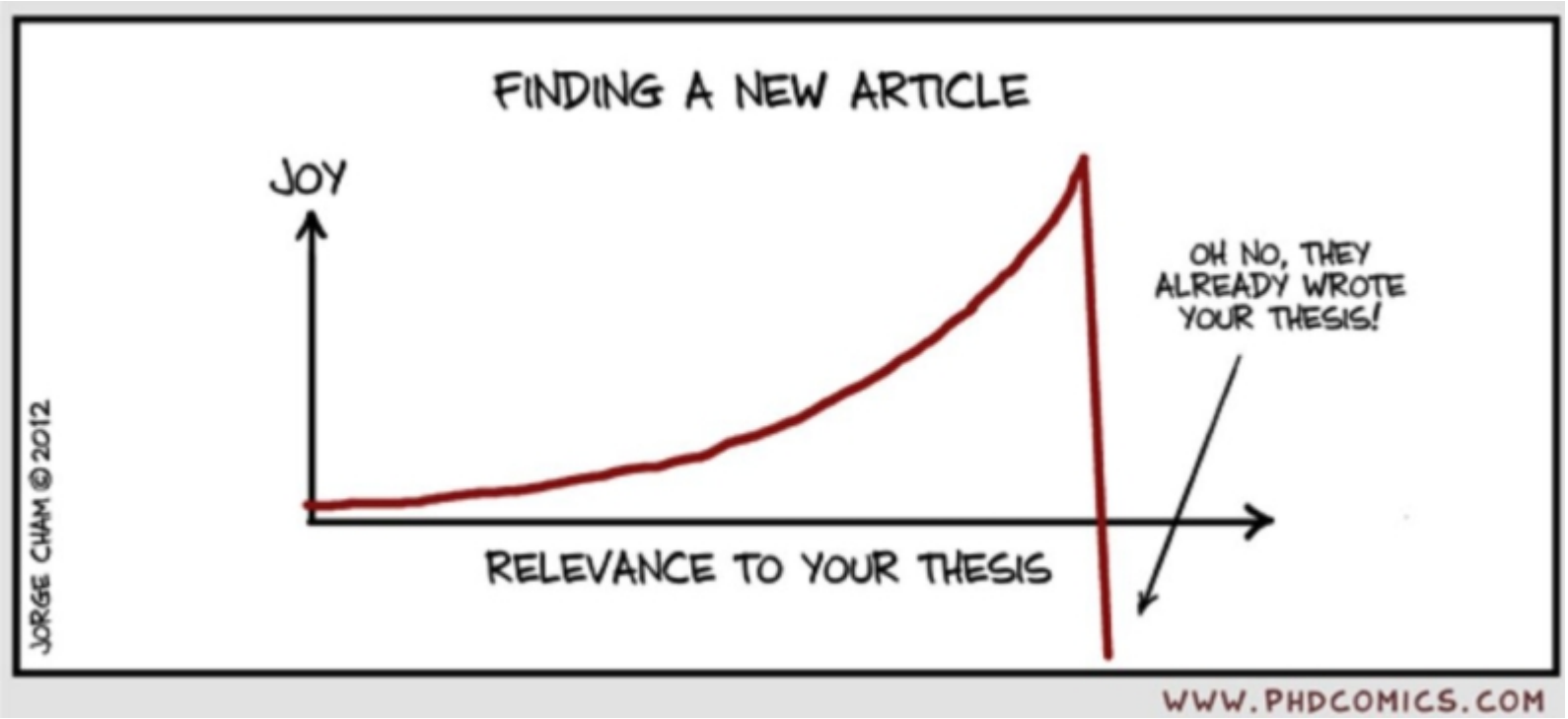


如何查找论文 （给定关键词）

维基百科
（搜索引擎）

中文综述
（CNKI）

英文论文
（Google Scholar）



善用Google Scholar

善用Google Scholar

善用Google Scholar

查阅学者学术信息、引用情况，也提供引用格式文件

Latent dirichlet allocation

DM Blei, AY Ng, MI Jordan - Journal of machine Learning research, 2003 - jmlr.org

Abstract We describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying ...

Cited by 15978 Related articles All 124 versions Import into BibTeX Cite Save Fewer

学会使用相关搜索命令

- Author: “DM Blei”
- AllInTitle:“Latent dirichlet allocation”
- …

如何判断论文是否值得阅读

- 作者是否大牛学者？作者机构是否顶尖？
- 是否发表在顶级期刊/会议上？
- 论文社会关注度如何？
- 是否获得最佳论文？引用情况如何？

学术资源 ACM & IEEE

- 美国计算机学会
- 全球最大的计算机学术组织
- ACM 拥有大量高水平论文

- 电气和电子工程师协会
- 全球最大的电子与信息科学协会

如何判断论文是否值得阅读

arXiv.org

- 预印本文库
- 未发表的论文，良莠不齐
- 建议关注顶级组织的相关论文

---

subscribe Zhiyuan Liu

1 message

Zhiyuan Liu <liuzy@tsinghua.edu.cn>
To: cs@arxiv.org

add CL
add LG
add NE

如何判断论文是否值得阅读

新时代的论文途径

微博、TWITTER

微信公众号、知乎、GITHUB

…… …… …… ……

如何判断论文是否值得阅读

王晋东，现就职于微软亚洲研究院，研究方向为迁移学习和机器学习等。他在国际权威会议ICDM、UbiComp等发表多篇文章。

他是**知乎等知识共享社区的机器学习达人**（知乎用户名：王晋东不在家）

他还在**Github**上发起建立了多个与机器学习相关的资源仓库，成立了超过120个高校和研究所参与的机器学习群，**热心于知识的共享**

阅读论文的顺序

- 题目（1）
- 摘要（2）
- 正文
  - 导论（3）、相关工作、自己工作（5）、实验结果（4）、结论
- 致谢
- 参考文献（6）
- 附录

题目->摘要->导论->实验结果->自己工作->参考文献，要将精读和泛读充分结合，可能一年阅读文章成百上千篇，但是有几十篇文章一定需要仔细阅读，反复阅读

如何找好问题

- 一流学者提出问题
- 二流学者解决问题
- 三流学者打补丁

为什么找问题更重要、更难?

- 牛人提出问题往往能影响整个**领域的发展方向**
- 解决问题往往是个技术活，能够后天培养（理论素养、编程能力、写作能力等），而提出问题需要：**站得高、看得远、嗅觉好、当机立断、不畏风险**

如何找问题?

Think

differently

满腹经纶者固然可敬，擅长推陈出新者更值得推崇

哪里热闹去哪里

哪里人少去哪里

*“It is not worth an intelligent man's time to be in the majority. **By definition**, there are already enough people to do that.”*

--- G. H. Hardy (1877-1947)

如何找好问题

- 博览群书，对整个领域有全貌的把握
- 熟知学术界动态，知道当前最热门的问题是什么
- 明察秋毫，富有远见，结合个人兴趣选择一个数年后变成热门的领域，并全力以赴去做

做好不被认可的准备

Ludwig Boltzmann
1844–1906

Ludwig Eduard Boltzmann was an Austrian physicist who created the field of statistical mechanics. Prior to Boltzmann, the concept of entropy was already known from classical thermodynamics where it quantifies the fact that when we take energy from a system, not all of that energy is typically available to do useful work. Boltzmann showed that the thermodynamic entropy S, a macroscopic quantity, could be related to the statistical properties at the microscopic level. This is expressed through the famous equation $S = k \ln W$ in which W represents the number of possible microstates in a macrostate, and $k \simeq 1.38 \times 10^{-23}$ (in units of Joules per Kelvin) is known as Boltzmann's constant. Boltzmann's ideas were disputed by many scientists of they day. One difficulty they saw arose from the second law of thermodynamics, which states that the entropy of a closed system tends to increase with time. By contrast, at the microscopic level the classical Newtonian equations of physics are reversible, and so they found it difficult to see how the latter could explain the former. They didn't fully appreciate Boltzmann's arguments, which were statistical in nature and which concluded not that entropy could never decrease over time but simply that with overwhelming probability it would generally increase. Boltzmann even had a long-running dispute with the editor of the leading German physics journal who refused to let him refer to atoms and molecules as anything other than convenient theoretical constructs. The continued attacks on his work lead to bouts of depression, and eventually he committed suicide. Shortly after Boltzmann's death, new experiments by Perrin on colloidal suspensions verified his theories and confirmed the value of the Boltzmann constant. The equation $S = k \ln W$ is carved on Boltzmann's tombstone.

做好不被认可的准备

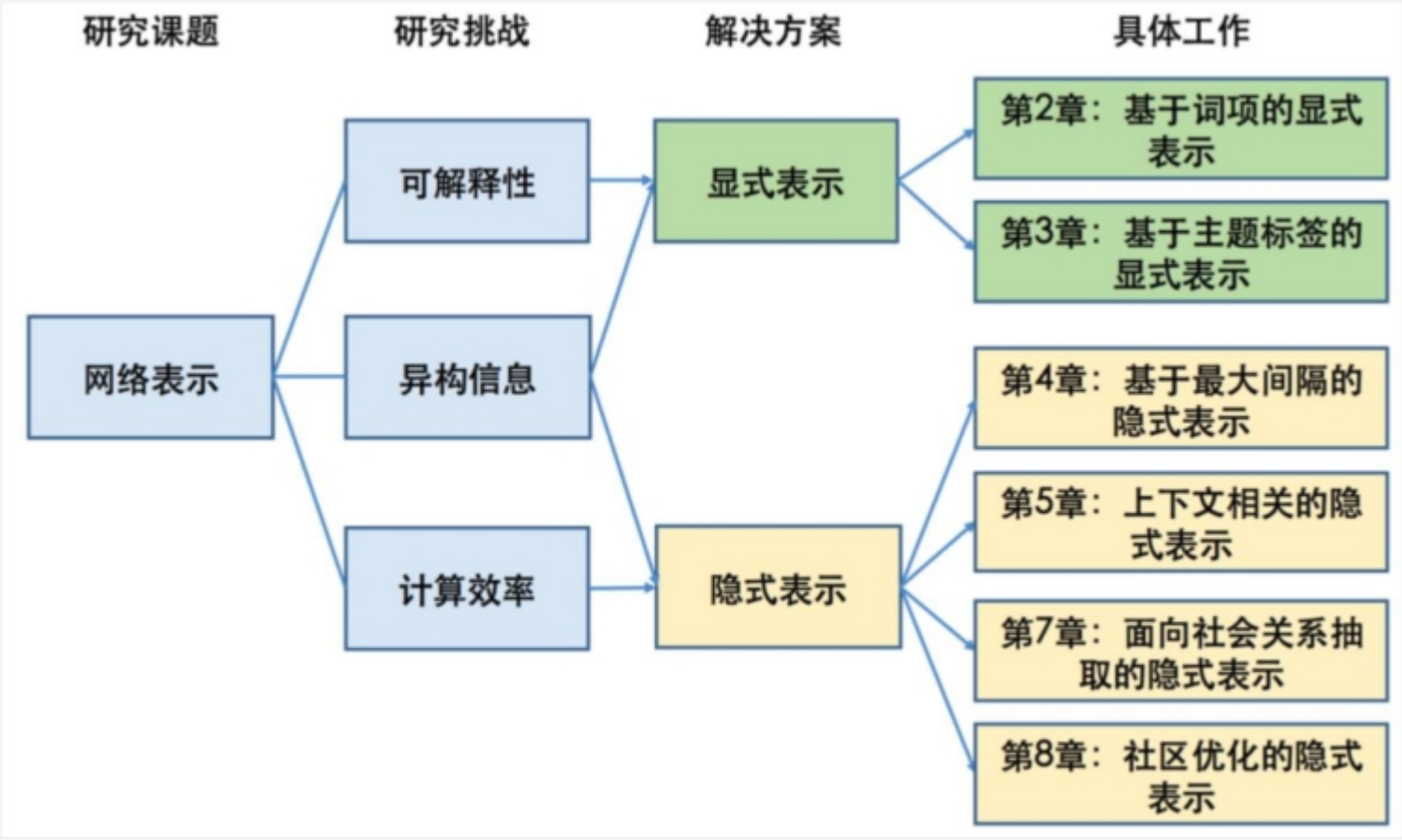
Frank Rosenblatt
1928–1969

Rosenblatt's perceptron played an important role in the history of machine learning. Initially, Rosenblatt simulated the perceptron on an IBM 704 computer at Cornell in 1957, but by the early 1960s he had built special-purpose hardware that provided a direct, parallel implementation of perceptron learning. Many of his ideas were encapsulated in "Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms" published in 1962. Rosenblatt's work was criticized by Marvin Minksy, whose objections were published in the book "Perceptrons", co-authored with Seymour Papert. This book was widely misinterpreted at the time as showing that neural networks were fatally flawed and could only learn solutions for linearly separable problems. In fact, it only proved such limitations in the case of single-layer networks such as the perceptron and merely conjectured (incorrectly) that they applied to more general network models. Unfortunately, however, this book contributed to the substantial decline in research funding for neural computing, a situation that was not reversed until the mid-1980s. Today, there are many hundreds, if not thousands, of applications of neural networks in widespread use, with examples in areas such as handwriting recognition and information retrieval being used routinely by millions of people.

对博士生的选题建议

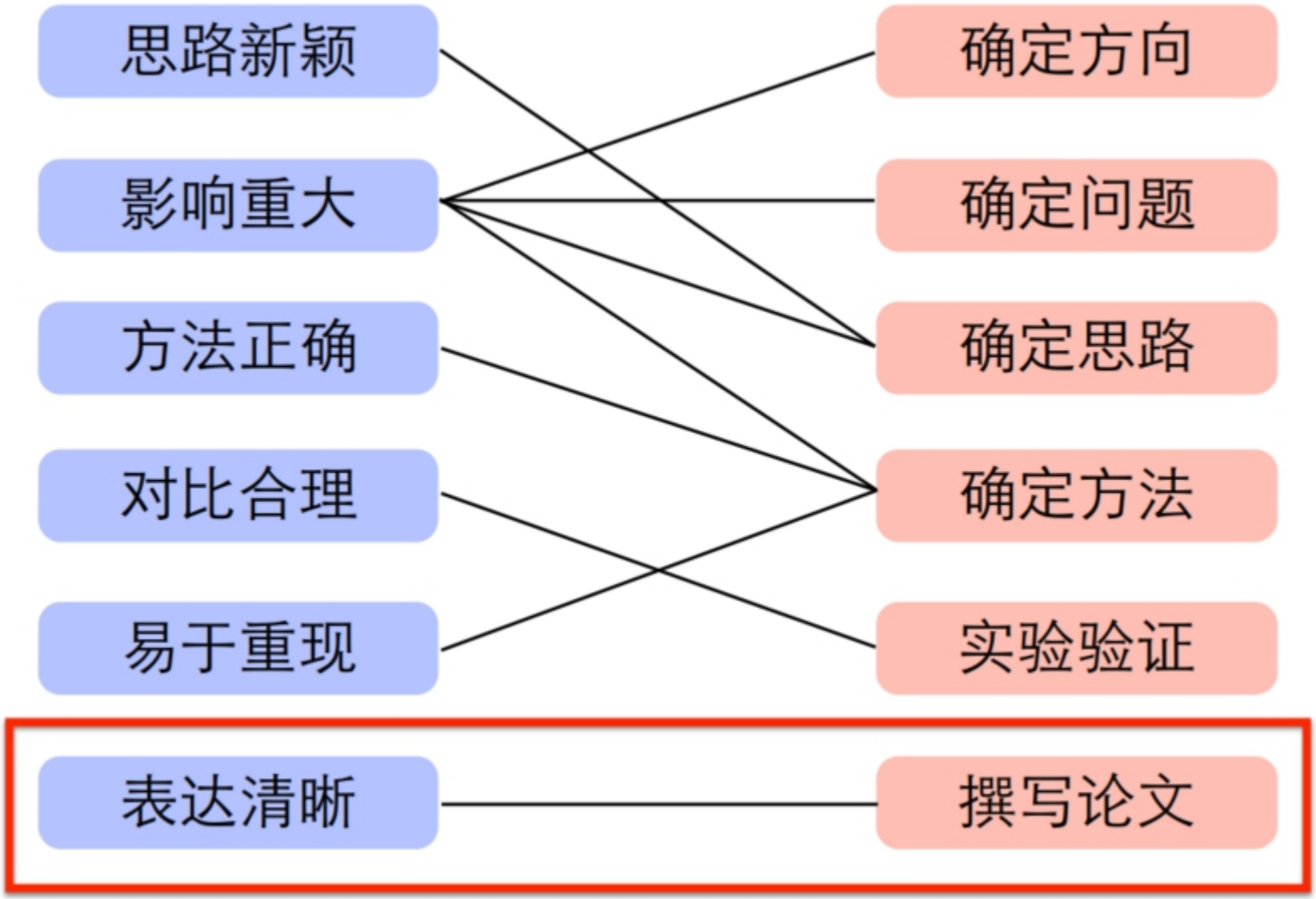
- 不只训练对单独一份工作选题的能力
- 思考博士生涯的整体选题（3-5个独立工作）

对博士生的选题建议



涂存超（2018）：面向社会计算的网络表示学习

写论文时什么最重要?



关于审稿

- 你以为审稿人应该是这样审稿的

审稿人一定是专家，无所不知。打印出来，仔细研读揣摩数天，对于看不懂的地方反复推敲。即使你的英文写得极其糟糕、即使你的文章组织很混乱、即使你的表述很难看懂，审稿人花费了大量的时间后终于看懂了，他认为你的工作是有意义的，决定给你个borderline或以上的分数。

关于审稿

- 你以为审稿人应该是这样审稿的

审稿人一定是专家，无所不知。打印出来，仔细研读揣摩数天，对于看不懂的地方反复推敲。即使你的英文写得极其糟糕、即使你的文章组织很混乱、即使你的表述很难看懂，审稿人花费了大量的时间后终于看懂了，他认为你的工作是有意义的，决定给你个border line或以上的分数。

- 实际上他们往往是这样的

他不一定是专家，一直忙于其他事，在deadline到来之前一天要完成n篇。审稿时他往往先看题目、摘要，扫一下introduction（知道你做什么），然后直接翻到最后找核心实验结果（做得好不好），然后基本确定录还是不录（也许只用5分钟！）。如果决定录，剩下就是写些赞美的话，指出些次要的小毛病。如果决定拒，下面的过程就是细看中间部分找理由拒了。

关于审稿

第一印象决定accept or reject

所以要 5 分钟以内打动审稿人

微博上的佐证

胡云华MSRA V　+加关注

最近有很多论文需要评审，跟同行聊天，得出一个有意思的结论：如果一篇论文在看完abstract和conclusion后还不能判断论文是否有价值的话，基本上这篇论文也就悲剧了。自己试了多次，屡试不爽。最极端的一篇我看了整整两天，全部搞懂作者在说什么后，仍然觉得应该拒掉，就跟只看5分钟得出的结论一致。

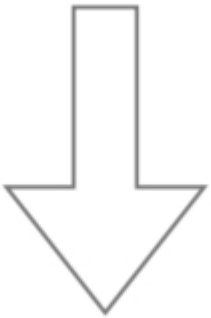
胡云华MSRA V：回复@shirlywang1983:我说的是"小论文"，毕业论文之类的评审得少，不好说。好的论文需要准确提炼观点，让读者在尽量短的时间内明白你做了什么，你的贡献是什么。如果自己没想清楚，肯定写不清楚的。当然这个过程很不容易，没有深厚积累谁都做不到。(12月5日 09:01)

kingdy9：说明第一印象很重要，也很准确。。有了第一印象后再找找文章中值得批判的地方就好了。。//@朱小燕THU:悲哀的是，已经感觉到了，但是为了写评语还是要看到底 (12月5日 09:38)

王伟DL：回复@胡云华MSRA:谢谢！我得修正我的观点，很同意"审论文时，abstract和conclusion写不好但内容好的情况少之又少。" (12月5日 14:22)

观念转变

以作者为核心整理工作



以读者为核心阐述工作

全心全意为读者服务

信息的呈现符合读者的认知惯性

深入浅出，引人入胜，让读者快速找到想要的信息

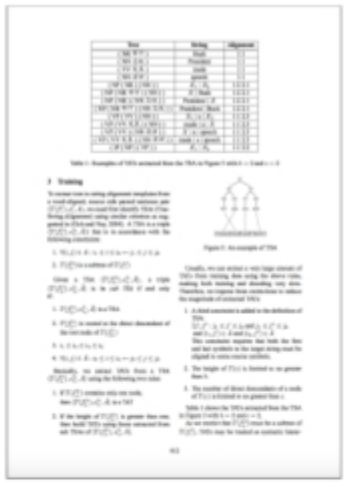
尽量降低读者的理解难度

合理地综合使用信息元素：图>曲线>表>正文>公式

尽量提高读者阅读时的愉悦感

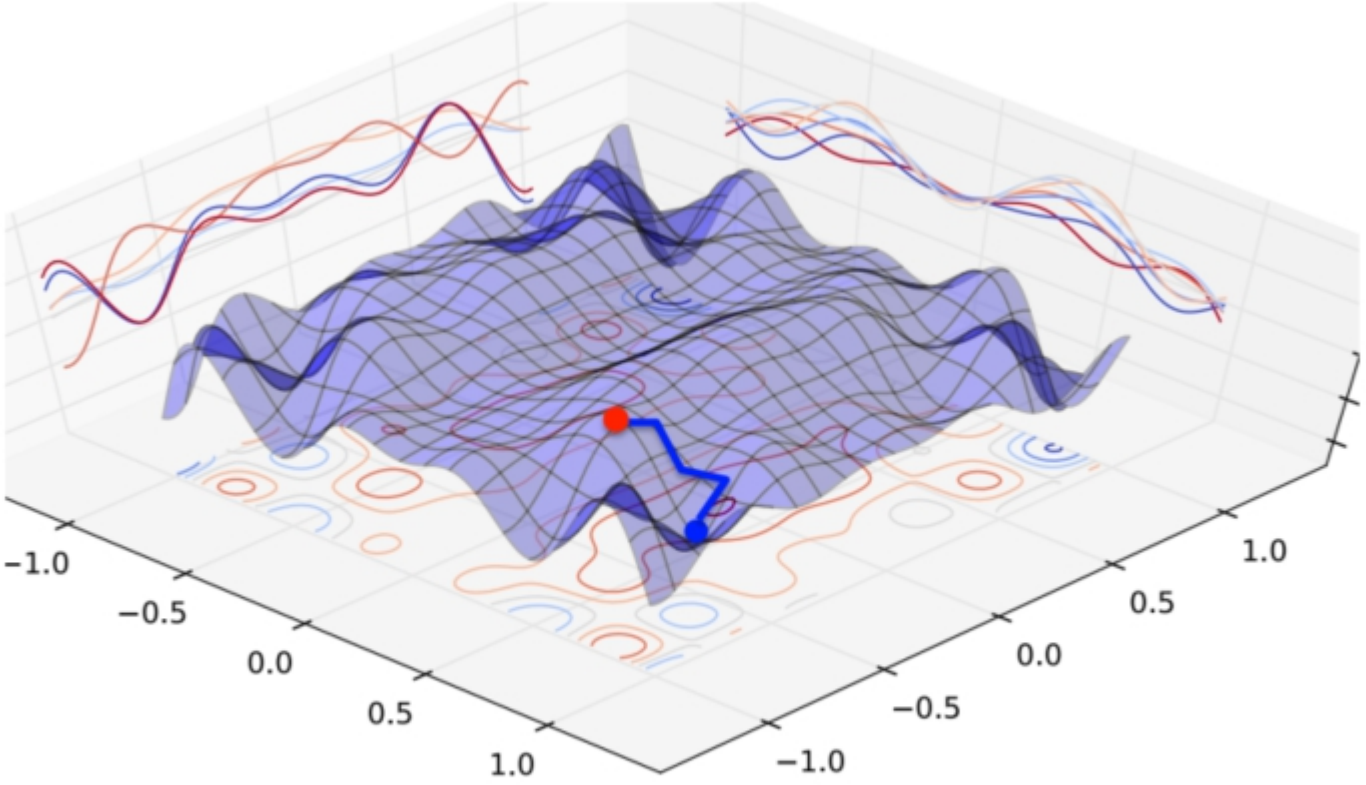
思想新颖、组织合理、逻辑严密
论证充分、文笔优美、排版美观

层次：论文



读者看到的是论文

层次：信息



阅读实际是信息接受的过程

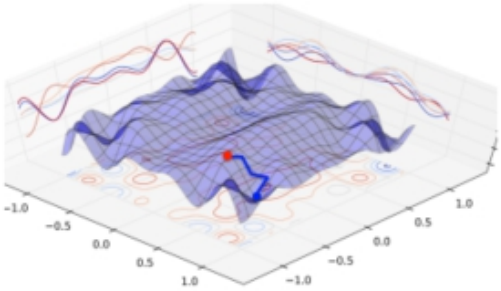
层次：思想

深层次反映的是作者的思想

阅读与写作

论文

信息



思想

阅读

写作

标题的重要性

- 如何看浩如烟海的文献?
  - 根据标题过滤50%
  - 根据摘要再过滤20%
  - 根据介绍再过滤20%
  - 剩下的10%再仔细看论文

黄铠

例子

Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data

John Lafferty[†*] LAFFERTY@CS.CMU.EDU
Andrew McCallum[*†] MCCALLUM@WHIZBANG.COM
Fernando Pereira[*‡] FPEREIRA@WHIZBANG.COM

[*]WhizBang! Labs–Research, 4616 Henry Street, Pittsburgh, PA 15213 USA
[†]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA
[‡]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104 USA

- 用一句话概括你所做的工作
- 考虑搜索引擎的影响，包含关键词

例子

Coooooooooooooollllllllllll!!!!!!!!!!!!!
Using Word Lengthening to Detect Sentiment in Microblogs

Samuel Brody
School of Communication
and Information
Rutgers University
sdbrody@gmail.com

Nicholas Diakopoulos
School of Communication
and Information
Rutgers University
diakop@rutgers.edu

- 可以适当地别出心裁

摘要

- 几句话概括你的工作
- 误区
  - 力图所有细节说清楚
  - 用很专业的术语描述
  - 出现数学符号

用语要简单，外行能看懂

摘要

问题是什么

我们大概怎么做的

我们做了什么

我们做得挺不错!

Abstract

Conventional n-best reranking techniques often suffer from the limited scope of the n-best list, which rules out many potentially good alternatives. We instead propose *forest reranking*, a method that reranks a packed forest of exponentially many parses. Since exact inference is intractable with non-local features, we present an approximate algorithm inspired by forest rescoring that makes discriminative training practical over the whole Treebank. Our final result, an F-score of 91.7, outperforms both 50-best and 100-best reranking baselines, and is better than any previously reported systems trained on the Treebank.

引言

- 比题目和摘要更进一步，几段话说清你的工作
- 要点是充分论证你工作的必要和重要性，要让审稿人认同并迫不及待的想看
- 行文逻辑严密，论证充分

引言

常见的逻辑：

- 说明问题是什么
- 简单罗列前人工作
- 描述我们的工作

更好的逻辑：

- 说明问题是什么
- 目前最好的工作面临什么挑战
- 我们的方法能缓解上述挑战

引言

Improving Tree-to-Tree Translation with Packed Forests

Yang Liu and **Yajuan Lü** and **Qun Liu**
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{yliu,lvyajuan,liuqun}@ict.ac.cn

Abstract

Current tree-to-tree models suffer from parsing errors as they usually use only 1-best parses for rule extraction and decoding. We instead propose a forest-based tree-to-tree model that uses packed forests. The model is based on a probabilistic synchronous tree substitution grammar (STSG), which can be learned from aligned forest pairs automatically. The decoder finds ways of decomposing trees in the source forest into elementary trees using the source projection of STSG while building target forest in parallel. Comparable to the state-of-the-art phrase-based system Moses, using packed forests in tree-to-tree translation results in a significant absolute improvement of 3.6 BLEU points over using 1-best trees.

1 Introduction

Approaches to syntax-based statistical machine translation make use of parallel data with syntactic annotations, either in the form of phrase structure trees or dependency trees. They can be roughly divided into three categories: *string-to-tree* models (e.g., (Galley et al., 2006; Marcu et al., 2006; Shen et al., 2008)), *tree-to-string* models (e.g., (Liu et al., 2006; Huang et al., 2006)), and *tree-to-tree* models (e.g., (Eisner, 2003; Ding and Palmer, 2005; Cowan et al., 2006; Zhang et al., 2008)). By modeling the syntax of both source and target languages, tree-to-tree approaches have the potential benefit of providing rules linguistically better motivated. However, while string-to-tree and tree-to-string models demonstrate promising results in empirical evaluations, tree-to-tree models have still been underachieving.

We believe that tree-to-tree models face two major challenges. First, tree-to-tree models are more vulnerable to parsing errors. Obtaining syntactic annotations in quantity usually entails running automatic parsers on a parallel corpus. As the amount and domain of the data used to train parsers are relatively limited, parsers will inevitably output ill-formed trees when handling real-world text. Guided by such noisy syntactic information, syntax-based models that rely on 1-best parses are prone to learn noisy translation rules in training phase and produce degenerate translations in decoding phase (Quirk and Corston-Oliver, 2006). This situation aggravates for tree-to-tree models that use syntax on both sides.

Second, tree-to-tree rules provide poorer rule coverage. As a tree-to-tree rule requires that there must be trees on both sides, tree-to-tree models lose a larger amount of linguistically unmotivated mappings. Studies reveal that the absence of such non-syntactic mappings will impair translation quality dramatically (Marcu et al., 2006; Liu et al., 2007; DeNeefe et al., 2007; Zhang et al., 2008).

Compactly encoding exponentially many parses, *packed forests* prove to be an excellent fit for alleviating the above two problems (Mi et al., 2008; Mi and Huang, 2008). In this paper, we propose a forest-based tree-to-tree model. To learn STSG rules from aligned forest pairs, we introduce a series of notions for identifying minimal tree-to-tree rules. Our decoder first converts the source forest to a translation forest and then finds the best derivation that has the source yield of one source tree in the forest. Comparable to Moses, our forest-based tree-to-tree model achieves an absolute improvement of 3.6 BLEU points over conventional tree-based model.

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 558–566,
Suntec, Singapore, 2-7 August 2009.

问题

Approaches to syntax-based statistical machine translation make use of parallel data with syntactic annotations, either in the form of phrase structure trees or dependency trees. They can be roughly divided into three categories: *string-to-tree* models (e.g., (Galley et al., 2006; Marcu et al., 2006; Shen et al., 2008)), *tree-to-string* models (e.g., (Liu et al., 2006; Huang et al., 2006)), and *tree-to-tree* models (e.g., (Eisner, 2003; Ding and Palmer, 2005; Cowan et al., 2006; Zhang et al., 2008)). By modeling the syntax of both source and target languages, tree-to-tree approaches have the potential benefit of providing rules linguistically better motivated. However, while string-to-tree and tree-to-string models demonstrate promising results in empirical evaluations, tree-to-tree models have still been underachieving.

引言

Improving Tree-to-Tree Translation with Packed Forests

Yang Liu and **Yajuan Lü** and **Qun Liu**
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{yliu,lvyajuan,liuqun}@ict.ac.cn

Abstract

Current tree-to-tree models suffer from parsing errors as they usually use only 1-best parses for rule extraction and decoding. We instead propose a forest-based tree-to-tree model that uses packed forests. The model is based on a probabilistic synchronous tree substitution grammar (STSG), which can be learned from aligned forest pairs automatically. The decoder finds ways of decomposing trees in the source forest into elementary trees using the source projection of STSG while building target forest in parallel. Comparable to the state-of-the-art phrase-based system Moses, using packed forests in tree-to-tree translation results in a significant absolute improvement of 3.6 BLEU points over using 1-best trees.

1 Introduction

Approaches to syntax-based statistical machine translation make use of parallel data with syntactic annotations, either in the form of phrase structure trees or dependency trees. They can be roughly divided into three categories: *string-to-tree* models (e.g., (Galley et al., 2006; Marcu et al., 2006; Shen et al., 2008)), *tree-to-string* models (e.g., (Liu et al., 2006; Huang et al., 2006)), and *tree-to-tree* models (e.g., (Eisner, 2003; Ding and Palmer, 2005; Cowan et al., 2006; Zhang et al., 2008)). By modeling the syntax of both source and target languages, tree-to-tree approaches have the potential benefit of providing rules linguistically better motivated. However, while string-to-tree and tree-to-string models demonstrate promising results in empirical evaluations, tree-to-tree models have still been underachieving.

We believe that tree-to-tree models face two major challenges. First, tree-to-tree models are more vulnerable to parsing errors. Obtaining syntactic annotations in quantity usually entails running automatic parsers on a parallel corpus. As the amount and domain of the data used to train parsers are relatively limited, parsers will inevitably output ill-formed trees when handling real-world text. Guided by such noisy syntactic information, syntax-based models that rely on 1-best parses are prone to learn noisy translation rules in training phase and produce degenerate translations in decoding phase (Quirk and Corston-Oliver, 2006). This situation aggravates for tree-to-tree models that use syntax on both sides.

Second, tree-to-tree *rules* provide poorer rule coverage. As a tree-to-tree rule requires that there must be trees on both sides, tree-to-tree models lose a larger amount of linguistically unmotivated mappings. Studies reveal that the absence of such non-syntactic mappings will impair translation quality dramatically (Marcu et al., 2006; Liu et al., 2007; DeNeefe et al., 2007; Zhang et al., 2008).

Compactly encoding exponentially many parses, *packed forests* prove to be an excellent fit for alleviating the above two problems (Mi et al., 2008; Mi and Huang, 2008). In this paper, we propose a forest-based tree-to-tree model. To learn STSG rules from aligned forest pairs, we introduce a series of notions for identifying minimal tree-to-tree rules. Our decoder first converts the source forest to a translation forest and then finds the best derivation that has the source yield of one source tree in the forest. Comparable to Moses, our forest-based tree-to-tree model achieves an absolute improvement of 3.6 BLEU points over conventional tree-based model.

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 558–566,
Suntec, Singapore, 2-7 August 2009.

挑战

We believe that tree-to-tree models face two major challenges. First, tree-to-tree models are more vulnerable to parsing errors. Obtaining syntactic annotations in quantity usually entails running automatic parsers on a parallel corpus. As the amount and domain of the data used to train parsers are relatively limited, parsers will inevitably output ill-formed trees when handling real-world text. Guided by such noisy syntactic information, syntax-based models that rely on 1-best parses are prone to learn noisy translation rules in training phase and produce degenerate translations in decoding phase (Quirk and Corston-Oliver, 2006). This situation aggravates for tree-to-tree models that use syntax on both sides.

Second, tree-to-tree rules provide poorer rule coverage. As a tree-to-tree rule requires that there must be trees on both sides, tree-to-tree models lose a larger amount of linguistically unmotivated mappings. Studies reveal that the absence of such non-syntactic mappings will impair translation quality dramatically (Marcu et al., 2006; Liu et al., 2007; DeNeefe et al., 2007; Zhang et al., 2008).

引言

Improving Tree-to-Tree Translation with Packed Forests

Yang Liu and **Yajuan Lü** and **Qun Liu**
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{yliu,lvyajuan,liuqun}@ict.ac.cn

Abstract

Current tree-to-tree models suffer from parsing errors as they usually use only 1-best parses for rule extraction and decoding. We instead propose a forest-based tree-to-tree model that uses packed forests. The model is based on a probabilistic synchronous tree substitution grammar (STSG), which can be learned from aligned forest pairs automatically. The decoder finds ways of decomposing trees in the source forest into elementary trees using the source projection of STSG while building target forest in parallel. Comparable to the state-of-the-art phrase-based system Moses, using packed forests in tree-to-tree translation results in a significant absolute improvement of 3.6 BLEU points over using 1-best trees.

1 Introduction

Approaches to syntax-based statistical machine translation make use of parallel data with syntactic annotations, either in the form of phrase structure trees or dependency trees. They can be roughly divided into three categories: *string-to-tree* models (e.g., (Galley et al., 2006; Marcu et al., 2006; Shen et al., 2008)), *tree-to-string* models (e.g., (Liu et al., 2006; Huang et al., 2006)), and *tree-to-tree* models (e.g., (Eisner, 2003; Ding and Palmer, 2005; Cowan et al., 2006; Zhang et al., 2008)). By modeling the syntax of both source and target languages, tree-to-tree approaches have the potential benefit of providing rules linguistically better motivated. However, while string-to-tree and tree-to-string models demonstrate promising results in empirical evaluations, tree-to-tree models have still been underachieving.

We believe that tree-to-tree models face two major challenges. First, tree-to-tree models are more vulnerable to parsing errors. Obtaining syntactic annotations in quantity usually entails running automatic parsers on a parallel corpus. As the amount and domain of the data used to train parsers are relatively limited, parsers will inevitably output ill-formed trees when handling real-world text. Guided by such noisy syntactic information, syntax-based models that rely on 1-best parses are prone to learn noisy translation rules in training phase and produce degenerate translations in decoding phase (Quirk and Corston-Oliver, 2006). This situation aggravates for tree-to-tree models that use syntax on both sides.

Second, tree-to-tree rules provide poorer rule coverage. As a tree-to-tree rule requires that there must be trees on both sides, tree-to-tree models lose a larger amount of linguistically unmotivated mappings. Studies reveal that the absence of such non-syntactic mappings will impair translation quality dramatically (Marcu et al., 2006; Liu et al., 2007; DeNeefe et al., 2007; Zhang et al., 2008).

Compactly encoding exponentially many parses, *packed forests* prove to be an excellent fit for alleviating the above two problems (Mi et al., 2008; Mi and Huang, 2008). In this paper, we propose a forest-based tree-to-tree model. To learn STSG rules from aligned forest pairs, we introduce a series of notions for identifying minimal tree-to-tree rules. Our decoder first converts the source forest to a translation forest and then finds the best derivation that has the source yield of one source tree in the forest. Comparable to Moses, our forest-based tree-to-tree model achieves an absolute improvement of 3.6 BLEU points over conventional tree-based model.

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 558–566,
Suntec, Singapore, 2-7 August 2009.

我们的工作

Compactly encoding exponentially many parses, *packed forests* prove to be an excellent fit for alleviating the above two problems (Mi et al., 2008; Mi and Huang, 2008). In this paper, we propose a forest-based tree-to-tree model. To learn STSG rules from aligned forest pairs, we introduce a series of notions for identifying minimal tree-to-tree rules. Our decoder first converts the source forest to a translation forest and then finds the best derivation that has the source yield of one source tree in the forest. Comparable to Moses, our forest-based tree-to-tree model achieves an absolute improvement of 3.6 BLEU points over conventional tree-based model.

引言

每个段落有个论断式的中心句

其余部分都是支撑句，围绕中心句展开论证

- 前人工作
- 具体数据

支撑句之间可分类组织

段尾可加上衔接句

引言

The need to segment and label sequences arises in many different problems in several scientific fields. Hidden Markov models (HMMs) and stochastic grammars are well understood and widely used probabilistic models for such problems. In computational biology, HMMs and stochastic grammars have been successfully used to align biological sequences, find sequences homologous to a known evolutionary family, and analyze RNA secondary structure (Durbin et al., 1998). In computational linguistics and computer science, HMMs and stochastic grammars have been applied to a wide variety of problems in text and speech processing, including topic segmentation, part-of-speech (POS) tagging, information extraction, and syntactic disambiguation (Manning & Schütze, 1999).

中心句与支撑句

引言

中心句与支撑句

We believe that it is important to make available to syntax-based models all the bilingual phrases that are typically available to phrase-based models. On one hand, phrases have been proven to be a simple and powerful mechanism for machine translation. They excel at capturing translations of short idioms, providing local re-ordering decisions, and incorporating context information straightforwardly. Chiang (2005) shows significant improvement by keeping the strengths of phrases while incorporating syntax into statistical translation. On the other hand, the performance of linguistically syntax-based models can be hindered by making use of only syntactic phrase pairs. Studies reveal that linguistically syntax-based models are sensitive to syntactic analysis (Quirk and Corston-Oliver, 2006), which is still not reliable enough to handle real-world texts due to limited size and domain of training data.

引言

Finding word alignments between parallel texts, however, is still far from a trivial work due to the diversity of natural languages. For example, the alignment of words within idiomatic expressions, free translations, and missing content or function words is problematic. When two languages widely differ in word order, finding word alignments is especially hard. Therefore, it is necessary to incorporate all useful linguistic information to alleviate these problems.

衔接句

Tiedemann (2003) introduced a word alignment approach based on combination of association clues. Clues combination is done by disjunction of single clues, which are defined as probabilities of associations. The crucial assumption of clue combination that clues are independent of each other, however, is not always true. Och and Ney (2003) proposed

引言

新技巧：

- 在首页放一个图或者表，让读者一目了然你的工作
- 不要写“This paper is organized as follows”，而是直接列出自己的贡献

引言

眼动仪的佐证：

图片来自清华大学刘奕群

读者潜意识里优先选择易理解度高的信息元素

리言

Forest Reranking: Discriminative Parsing with Non-Local Features*

Liang Huang
University of Pennsylvania
Philadelphia, PA 19104
lhuang3@cis.upenn.edu

Abstract

Conventional n-best reranking techniques often suffer from the limited scope of the n-best list, which rules out many potentially good alternatives. We instead propose *forest reranking*, a method that reranks a packed forest of exponentially many parses. Since exact inference is intractable with non-local features, we present an approximate algorithm inspired by forest rescoring that makes discriminative training practical over the whole Treebank. Our final result, an F-score of 91.7, outperforms both 50-best and 100-best reranking baselines, and is better than any previously reported systems trained on the Treebank.

1 Introduction

Discriminative reranking has become a popular technique for many NLP problems, in particular, parsing (Collins, 2000) and machine translation (Shen et al., 2005). Typically, this method first generates a list of top-n candidates from a baseline system, and then reranks this n-best list with arbitrary features that are not computable or intractable to

	local	*non-local*
conventional reranking	only at the root	
DP-based discrim. parsing	exact	N/A
this work: forest-reranking	exact	*on-the-fly*

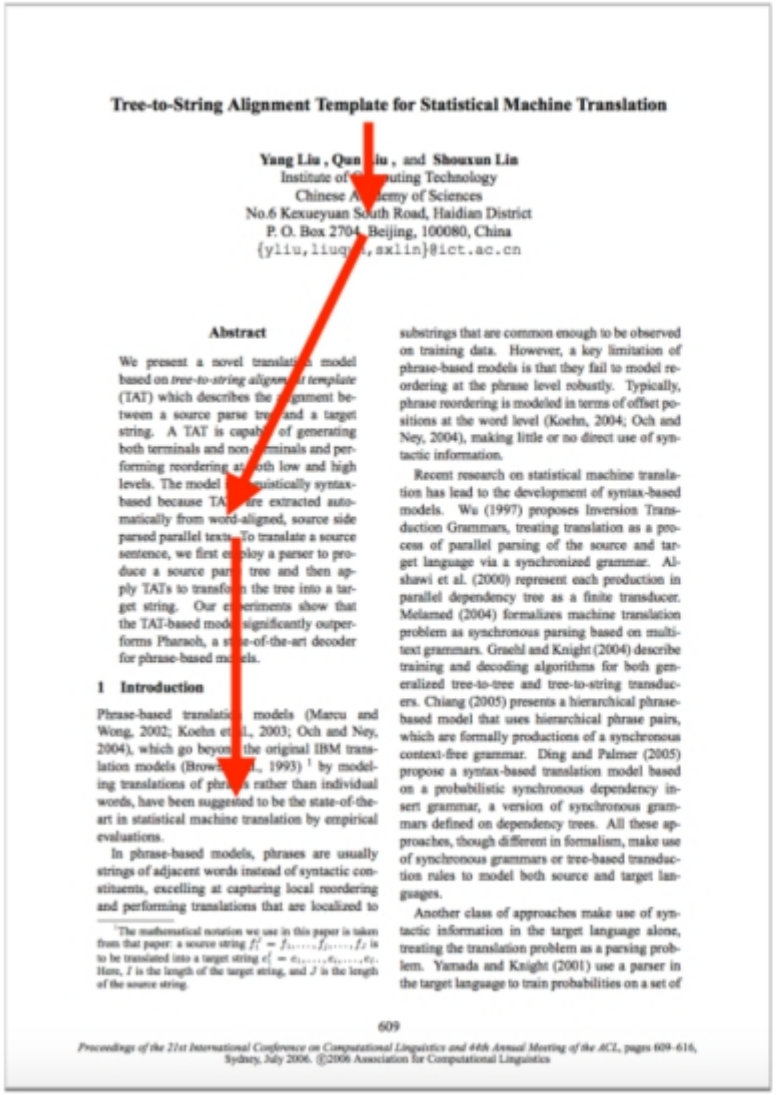
Table 1: Comparison of various approaches for incorporating local and non-local features.

sentence length. As a result, we often see very few variations among the n-best trees, for example, 50-best trees typically just represent a combination of 5 to 6 binary ambiguities (since $2^5 < 50 < 2^6$).

Alternatively, discriminative parsing is tractable with exact and efficient search based on dynamic programming (DP) if all features are restricted to be *local*, that is, only looking at a local window within the factored search space (Taskar et al., 2004; McDonald et al., 2005). However, we miss the benefits of non-local features that are not representable here.

Ideally, we would wish to combine the merits of both approaches, where an efficient inference algorithm could integrate both local and non-local features. Unfortunately, exact search is intractable (at least in theory) for features with unbounded scope.

引言 （信息流的变化）



Tree-to-String Alignment Template for Statistical Machine Translation

Yang Liu , Qun Liu , and Shouxun Lin
Institute of Computing Technology
Chinese Academy of Sciences
No.6 Kexueyuan South Road, Haidian District
P. O. Box 2704, Beijing, 100080, China
{yliu,liuqun,sxlin}@ict.ac.cn

Abstract

We present a novel translation model based on *tree-to-string alignment template* (TAT) which describes the alignment between a source parse tree and a target string. A TAT is capable of generating both terminals and non-terminals and performing reordering at both low and high levels. The model is linguistically syntax-based because TATs are extracted automatically from word-aligned, source side parsed parallel texts. To translate a source sentence, we first employ a parser to produce a source parse tree and then apply TATs to transform the tree into a target string. Our experiments show that the TAT-based model significantly outperforms Pharaoh, a state-of-the-art decoder for phrase-based models.

1 Introduction

Phrase-based translation models (Marcu and Wong, 2002; Koehn et al., 2003; Och and Ney, 2004), which go beyond the original IBM translation models (Brown et al., 1993) [1] by modeling translations of phrases rather than individual words, have been suggested to be the state-of-the-art in statistical machine translation by empirical evaluations.

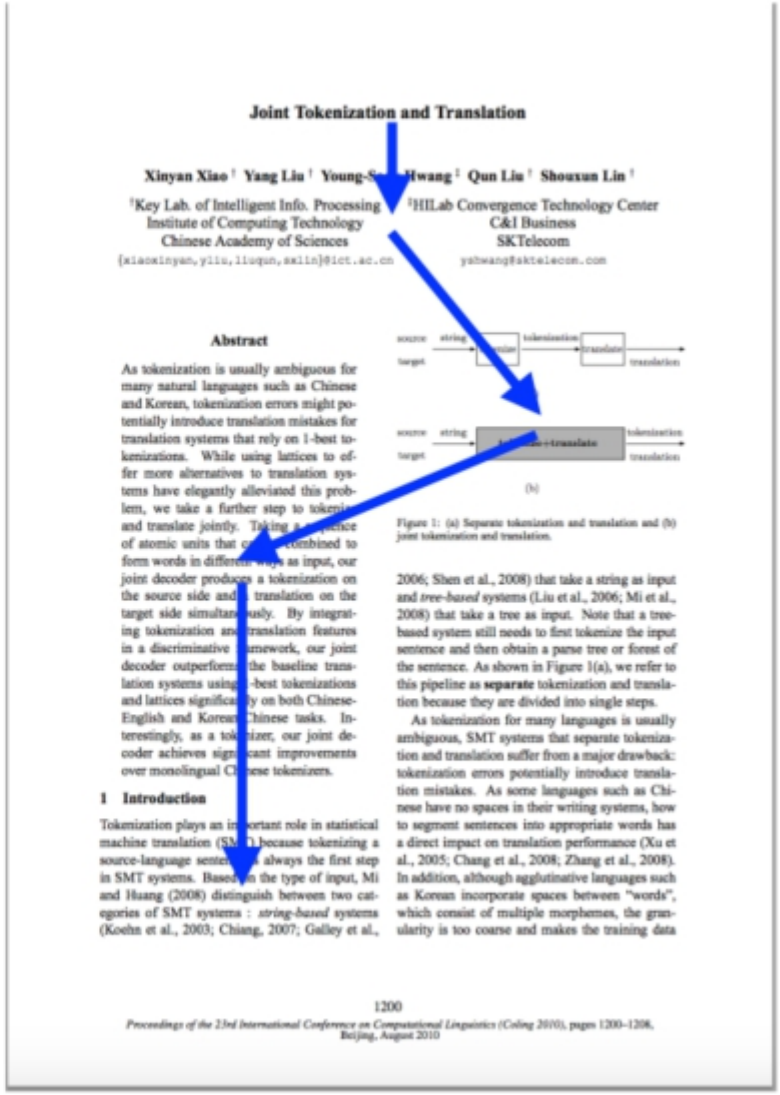
In phrase-based models, phrases are usually strings of adjacent words instead of syntactic constituents, excelling at capturing local reordering and performing translations that are localized to substrings that are common enough to be observed on training data. However, a key limitation of phrase-based models is that they fail to model reordering at the phrase level robustly. Typically, phrase reordering is modeled in terms of offset positions at the word level (Koehn, 2004; Och and Ney, 2004), making little or no direct use of syntactic information.

Recent research on statistical machine translation has lead to the development of syntax-based models. Wu (1997) proposes Inversion Transduction Grammars, treating translation as a process of parallel parsing of the source and target language via a synchronized grammar. Alshawi et al. (2000) represent each production in parallel dependency tree as a finite transducer. Melamed (2004) formalizes machine translation problem as synchronous parsing based on multi-text grammars. Graehl and Knight (2004) describe training and decoding algorithms for both generalized tree-to-tree and tree-to-string transducers. Chiang (2005) presents a hierarchical phrase-based model that uses hierarchical phrase pairs, which are formally productions of a synchronous context-free grammar. Ding and Palmer (2005) propose a syntax-based translation model based on a probabilistic synchronous dependency insert grammar, a version of synchronous grammars defined on dependency trees. All these approaches, though different in formalism, make use of synchronous grammars or tree-based transduction rules to model both source and target languages.

Another class of approaches make use of syntactic information in the target language alone, treating the translation problem as a parsing problem. Yamada and Knight (2001) use a parser in the target language to train probabilities on a set of

[1] The mathematical notation we use in this paper is taken from that paper: a source string $f_1^J = f_1, \ldots, f_j, \ldots, f_J$ is to be translated into a target string $e_1^I = e_1, \ldots, e_i, \ldots, e_I$. Here, I is the length of the target string, and J is the length of the source string.

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 609–616, Sydney, July 2006.



Joint Tokenization and Translation

Xinyan Xiao [†] Yang Liu [†] Young-Sook Hwang [‡] Qun Liu [†] Shouxun Lin [†]
[†]Key Lab. of Intelligent Info. Processing
Institute of Computing Technology
Chinese Academy of Sciences
{xiaoxinyan,yliu,liuqun,sxlin}@ict.ac.cn
[‡]HLLab Convergence Technology Center
C&I Business
SKTelecom
yshwang@sktelecom.com

Abstract

As tokenization is usually ambiguous for many natural languages such as Chinese and Korean, tokenization errors might potentially introduce translation mistakes for translation systems that rely on 1-best tokenizations. While using lattices to offer more alternatives to translation systems have elegantly alleviated this problem, we take a further step to tokenize and translate jointly. Taking a lattice of atomic units that can be combined to form words in different ways as input, our joint decoder produces a tokenization on the source side and a translation on the target side simultaneously. By integrating tokenization and translation features in a discriminative framework, our joint decoder outperforms the baseline translation systems using 1-best tokenizations and lattices significantly on both Chinese-English and Korean-Chinese tasks. Interestingly, as a tokenizer, our joint decoder achieves significant improvements over monolingual Chinese tokenizers.

1 Introduction

Tokenization plays an important role in statistical machine translation (SMT) because tokenizing a source-language sentence is always the first step in SMT systems. Based on the type of input, Mi and Huang (2008) distinguish between two categories of SMT systems : *string-based* systems (Koehn et al., 2003; Chiang, 2007; Galley et al., 2006; Shen et al., 2008) that take a string as input and *tree-based* systems (Liu et al., 2006; Mi et al., 2008) that take a tree as input. Note that a tree-based system still needs to first tokenize the input sentence and then obtain a parse tree or forest of the sentence. As shown in Figure 1(a), we refer to this pipeline as **separate** tokenization and translation because they are divided into single steps.

Figure 1: (a) Separate tokenization and translation and (b) joint tokenization and translation.

As tokenization for many languages is usually ambiguous, SMT systems that separate tokenization and translation suffer from a major drawback: tokenization errors potentially introduce translation mistakes. As some languages such as Chinese have no spaces in their writing systems, how to segment sentences into appropriate words has a direct impact on translation performance (Xu et al., 2005; Chang et al., 2008; Zhang et al., 2008). In addition, although agglutinative languages such as Korean incorporate spaces between "words", which consist of multiple morphemes, the granularity is too coarse and makes the training data

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1200–1208, Beijing, August 2010

引言 （图和表的重要性）

- 图和表是论文的骨架，争取让读者按照顺序看就能理解论文的主要思路，不用通过读正文
  - 一般第一遍看，都会看图、找例子
  - 然后到后边看主要结果
  - 再从头看正文

- 把论文的元素放在最应用放在的地方，符合读者的认知惯性，降低理解程度

引言 （直接列出自己的贡献）

coding phase. [1] Based on max-translation decoding and max-derivation decoding used in conventional *individual* decoders (Section 2), we go further to develop a *joint* decoder that integrates multiple models on a firm basis:

- Structuring the search space of each model as a *translation hypergraph* (Section 3.1), our joint decoder packs individual translation hypergraphs together by merging nodes that have identical partial translations (Section 3.2). Although such *translation-level combination* will not produce new translations, it does change the way of selecting promising candidates.
- Two models could even share derivations with each other if they produce the same structures on the target side (Section 3.3), which we refer to as *derivation-level combination*. This method enlarges the search space by allowing for mixing different types of translation rules within one derivation.
- As multiple derivations are used for finding optimal translations, we extend the minimum error rate training (MERT) algorithm (Och, 2003) to tune feature weights with respect to BLEU score for max-translation decoding (Section 4).

引言 （全局连贯性）

Methodology

如何描述自己的方法?

不要上来就描述工作，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中论文做理详细的解释
- 有利于对比Baseline和你的方法

Methodology

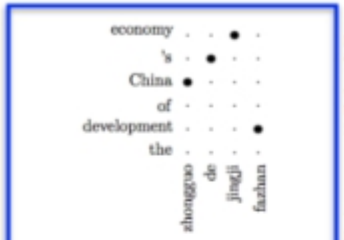


Figure 1: An example of word alignment between a pair of Chinese and English sentences.

ear space to encode the probabilities of exponentially many alignments. We develop a new algorithm for extracting phrase pairs from weighted matrices and show how to estimate their relative frequencies and lexical weights. Experimental results show that using weighted matrices achieves consistent improvements in translation quality and significant reduction in extraction time over using n-best lists.

2 Background

Figure 1 shows an example of word alignment between a pair of Chinese and English sentences. The Chinese and English words are listed horizontally and vertically, respectively. The dark points indicate the correspondence between the words in two languages. For example, the first Chinese word "*zhongguo*" is aligned to the fourth English word "*China*".

Formally, given a source sentence $\mathbf{f} = f_1^J = f_1, \ldots, f_j, \ldots, f_J$ and a target sentence $\mathbf{e} = e_1^I = e_1, \ldots, e_i, \ldots, e_I$, we define a link $l = (j, i)$ to exist if f_j and e_i are translation (or part of translation) of one another. Then, an alignment $\mathbf{a}$ is a subset of the Cartesian product of word positions:

$$\mathbf{a} \subseteq \{(j, i) : j = 1, \ldots, J; i = 1, \ldots, I\} \quad (1)$$

Usually, SMT systems only use the 1-best alignments for extracting translation rules. For example, given a source phrase $\bar{f}$ and a target phrase $\bar{e}$, the phrase pair $(\bar{f}, \bar{e})$ is said to be *consistent* (Och and Ney, 2004) with the alignment if and only if: (1) there must be at least one word inside one phrase aligned to a word inside the other phrase and (2) no words inside one phrase can be aligned to a word outside the other phrase.

After all phrase pairs are extracted from the training corpus, their translation probabilities can be estimated as *relative frequencies* (Och and Ney, 2004):

$$\phi(\bar{e}|\bar{f}) = \frac{count(\bar{f}, \bar{e})}{\sum_{\bar{e}'} count(\bar{f}, \bar{e}')} \quad (2)$$

where $count(\bar{f}, \bar{e})$ indicates how often the phrase pair $(\bar{f}, \bar{e})$ occurs in the training corpus.

Besides relative frequencies, *lexical weights* (Koehn et al., 2003) are widely used to estimate how well the words in $\bar{f}$ translate the words in $\bar{e}$. To do this, one needs first to estimate a lexical translation probability distribution $w(e|f)$ by relative frequency from the same word alignments in the training corpus:

$$w(e|f) = \frac{count(f, e)}{\sum_{e'} count(f, e')} \quad (3)$$

Note that a special source NULL token is added to each source sentence and aligned to each unaligned target word.

As the alignment $\hat{a}$ between a phrase pair $(\bar{f}, \bar{e})$ is retained during extraction, the lexical weight can be calculated as

$$p_w(\bar{e}|\bar{f}, \hat{a}) = \prod_{i=1}^{|\bar{e}|} \frac{1}{|\{j|(j, i) \in \hat{a}\}|} \sum w(e_i|f_j) \quad (4)$$

If there are multiple alignments $\hat{a}$ for a phrase pair $(\bar{f}, \bar{e})$, Koehn et al. (2003) choose the one with the highest lexical weight:

$$p_w(\bar{e}|\bar{f}) = \max_{\hat{a}} \left\{ p_w(\bar{e}|\bar{f}, \hat{a}) \right\} \quad (5)$$

Simple and effective, relative frequencies and lexical weights have become the standard features in modern discriminative SMT systems.

3 Weighted Alignment Matrix

We believe that offering more candidate alignments to extracting translation rules might help improve translation quality. Instead of using n-best lists (Venugopal et al., 2008), we propose a new structure called *weighted alignment matrix*.

We use an example to illustrate our idea. Figure 2(a) and Figure 2(b) show two alignments of a Chinese-English sentence pair. We observe that some links (e.g., (1,4) corresponding to the word

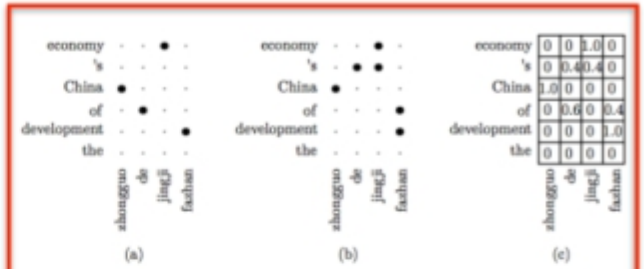


Figure 2: (a) One alignment of a sentence pair; (b) another alignment of the same sentence pair; (c) the resulting weighted alignment matrix that takes the two alignments as samples, of which the initial probabilities are 0.6 and 0.4, respectively.

pair ("*zhongguo*", "*China*")) occur in both alignments, some links (e.g., (2,3) corresponding to the word pair ("*de*","*of*")) occur only in one alignment, and some links (e.g., (1,1) corresponding to the word pair ("*zhongguo*", "*the*")) do not occur. Intuitively, we can estimate how well two words are aligned by calculating its relative frequency, which is the probability sum of alignments in which the link occurs divided by the probability sum of all possible alignments. Suppose that the probabilities of the two alignments in Figures 2(a) and 2(b) are 0.6 and 0.4, respectively. We can estimate the relative frequencies for every word pair and obtain a weighted matrix shown in Figure 2(c). Therefore, each word pair is associated with a probability to indicate how well they are aligned. For example, in Figure 2(c), we say that the word pair ("*zhongguo*", "*China*") is definitely aligned, ("*zhongguo*", "the") is definitely unaligned, and ("*de*", "*of*") has a 60% chance to get aligned.

Formally, a weighted alignment matrix m is a $J \times I$ matrix, in which each element stores a *link probability* $p_m(j, i)$ to indicate how well f_j and e_i are aligned. Currently, we estimate link probabilities from an n-best list by calculating relative frequencies:

$$p_m(j, i) = \frac{\sum_{a \in \mathcal{N}} p(a) \times \delta(a, j, i)}{\sum_{a \in \mathcal{N}} p(a)} \quad (6)$$

$$= \sum_{a \in \mathcal{N}} p(a) \times \delta(a, j, i) \quad (7)$$

where

$$\delta(a, j, i) = \begin{cases} 1 & (j, i) \in a \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

Note that $\mathcal{N}$ is an n-best list, $p(a)$ is the probability of an alignment a in the n-best list, $\delta(a, j, i)$ indicates whether a link (j, i) occurs in the alignment a or not. We assign 0 to any unseen alignment. As $p(a)$ is usually normalized (i.e., $\sum_{a \in \mathcal{N}} p(a) = 1$), we remove the denominator in Eq. (6).

Accordingly, the probability that the two words f_j and e_i are not aligned is

$$\bar{p}_m(j, i) = 1.0 - p_m(j, i) \quad (9)$$

For example, as shown in Figure 2(c), the probability for the two words "*de*" and "*of*" being aligned is 0.6 and the probability that they are not aligned is 0.4.

Intuitively, the probability of an alignment a is the product of link probabilities. If a link (j, i) occurs in a, we use $p_m(j, i)$; otherwise we use $\bar{p}_m(j, i)$. Formally, given a weighted alignment matrix m, the probability of an alignment a can be calculated as

$$p_m(a) = \prod_{j=1}^{J} \prod_{i=1}^{I} (p_m(j, i) \times \delta(a, j, i) + \bar{p}_m(j, i) \times (1 - \delta(a, j, i))) \quad (10)$$

It proves that the sum of all alignment probabilities is always 1: $\sum_{a \in \mathcal{A}} p_m(a) = 1$, where $\mathcal{A}$

Methodology （例子是利器）

英语不好说？用例子！

- 全篇使用一个running example，用来描述方法
- 围绕running example，展开描述工作
- 审稿人能够从中更舒服地了解你的工作，读正文会花费他更多的时间
- 看完running example，审稿人便知道核心思路

Methodology （逻辑顺序）

- 错误的顺序
  - 形式化描述
  - 解释数学符号的意义

- 正确的顺序
  - 首先给出running example
  - 然后利用running example，用通俗语言描述你的想法
  - 最后才是形式化描述

Methodology （逻辑顺序）

We believe that offering more candidate alignments to extracting translation rules might help improve translation quality. Instead of using n-best lists (Venugopal et al., 2008), we propose a new structure called *weighted alignment matrix*.

We use an example to illustrate our idea. Figure 2(a) and Figure 2(b) show two alignments of a Chinese-English sentence pair. We observe that some links (e.g., (1,4) corresponding to the word pair ("*zhongguo*", "*China*")) occur in both alignments, some links (e.g., (2,3) corresponding to the word pair ("*de*","*of*")) occur only in one alignment, and some links (e.g., (1,1) corresponding to the word pair ("*zhongguo*", "*the*")) do not occur. Intuitively, we can estimate how well two words are aligned by calculating its relative frequency, which is the probability sum of alignments in which the link occurs divided by the probability sum of all possible alignments. Suppose that the probabilities of the two alignments in Figures 2(a) and 2(b) are 0.6 and 0.4, respectively. We can estimate the relative frequencies for every word pair and obtain a weighted matrix shown in Figure 2(c). Therefore, each word pair is associated with a probability to indicate how well they are aligned. For example, in Figure 2(c), we say that the word pair ("*zhongguo*", "*China*") is definitely aligned, ("*zhongguo*", "the") is definitely unaligned, and ("*de*", "*of*") has a 60% chance to get aligned.

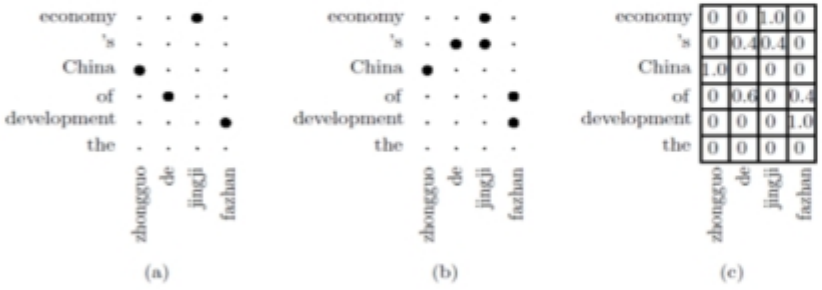


Figure 2: (a) One alignment of a sentence pair; (b) another alignment of the same sentence pair; (c) the resulting weighted alignment matrix that takes the two alignments as samples, of which the initial probabilities are 0.6 and 0.4, respectively.

Formally, a weighted alignment matrix m is a $J \times I$ matrix, in which each element stores a *link probability* $p_m(j,i)$ to indicate how well f_j and e_i are aligned. Currently, we estimate link probabilities from an n-best list by calculating relative frequencies:

Experiment

公认的标准数据和state-of-the-art系统

实验先辅后主

- 辅助实验：参数的影响
- 主实验：证明显著超过baseline

不辞劳苦，做到极致

Experiment

We first used the validation sets to find the optimal setting of our approach: noisy generation, the value of n, feature group, and training corpus size.

Table 2 shows the results of different noise generation strategies: randomly shuffling, inserting, replacing, and deleting words. We find shuffling source and target words randomly consistently yields the best results. One possible reason is that the translation probability product feature (Liu, Liu, and Lin, 2010) derived from GIZA++ suffices to evaluate lexical choices accurately. It is more important to guide the aligner to model the structural divergence by changing word orders randomly.

Table 3 gives the results of different values of sample size n on the validation sets. We find that increasing n does not lead to significant improvements. This might result from the high concentration property of log-linear models. Therefore, we simply set $n = 1$ in the following experiments.

Table 4 shows the effect of adding non-local features. As most structural divergence between natural languages are non-local, including non-local features leads to significant improvements for both French-English and Chinese-English. As a result, we used all 16 features in the following experiments.

Table 5 gives our final result on the test sets. Our approach outperforms all unsupervised aligners significantly statistically ($p < 0.01$) except for the Berkeley aligner on the French-English data. The margins on Chinese-English are generally much larger than French-English because Chinese and English are distantly related and exhibit more non-local structural divergence. Vigne used the same features as our system but was trained in a supervised way. Its results can be treated as the upper bounds that our method can potentially approach.

Experiment （caption信息丰富）

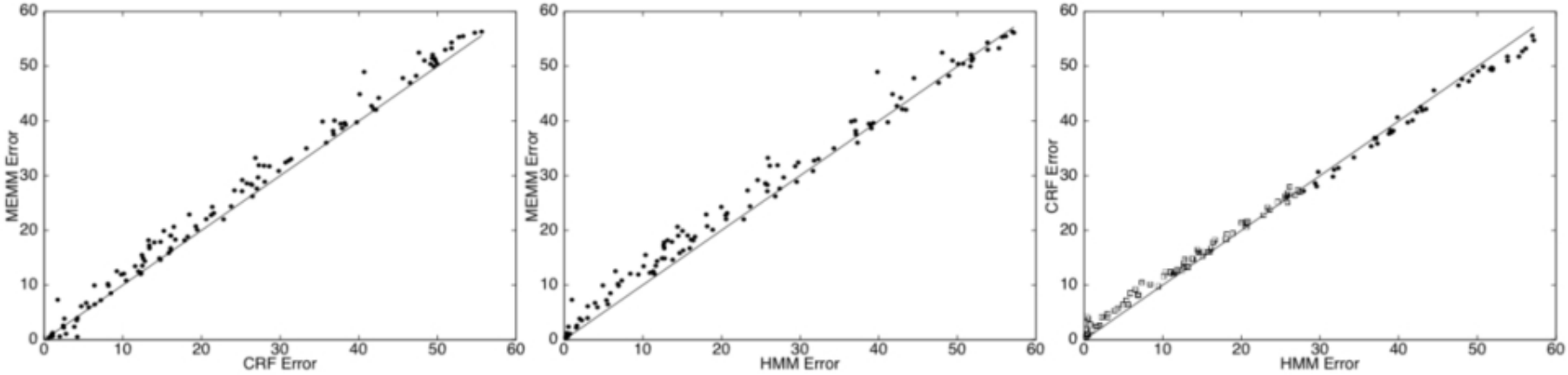


Figure 3. Plots of 2×2 error rates for HMMs, CRFs, and MEMMs on randomly generated synthetic data sets, as described in Section 5.2. As the data becomes "more second order," the error rates of the test models increase. As shown in the left plot, the CRF typically significantly outperforms the MEMM. The center plot shows that the HMM outperforms the MEMM. In the right plot, each open square represents a data set with $\alpha < \frac{1}{2}$, and a solid circle indicates a data set with $\alpha \geq \frac{1}{2}$. The plot shows that when the data is mostly second order ($\alpha \geq \frac{1}{2}$), the discriminatively trained CRF typically outperforms the HMM. These experiments are not designed to demonstrate the advantages of the additional representational power of CRFs and MEMMs relative to HMMs.

最好能直接看懂图，不用再去看正文

Related work

错误

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

通过批评乃至攻击前人工作证明你的工作的创新性

Related work

错误

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

通过批评乃至攻击前人工作证明你的工作的创新性

正确

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

Related work

2 Related Work

The CVG is inspired by two lines of research: Enriching PCFG parsers through more diverse sets of discrete states and recursive deep learning models that jointly learn classifiers and continuous feature representations for variable-sized inputs.

Improving Discrete Syntactic Representations

As mentioned in the introduction, there are several approaches to improving discrete representations for parsing. Klein and Manning (2003a) use manual feature engineering, while Petrov et al. (2006) use a learning algorithm that splits and merges the syntactic categories in order to maximize likelihood on the treebank. Their approach splits categories into several dozen subcategories. Another approach is lexicalized parsers (Collins, 2003; Charniak, 2000) that describe each category with a lexical item, usually the head word. More recently, Hall and Klein

Deep Learning and Recursive Deep Learning

Early attempts at using neural networks to describe phrases include Elman (1991), who used recurrent neural networks to create representations of sentences from a simple toy grammar and to analyze the linguistic expressiveness of the resulting representations. Words were represented as one-on vectors, which was feasible since the grammar only included a handful of words. Collobert and Weston (2008) showed that neural networks can perform well on sequence labeling lan-

Related work （传承与创新）

in a factored parser. We extend the above ideas from discrete representations to richer continuous ones. The CVG can be seen as factoring discrete and continuous parsing in one model. Another different approach to the above generative models is to learn discriminative parsers using many well designed features (Taskar et al., 2004; Finkel et al., 2008). We also borrow ideas from this line of research in that our parser combines the generative PCFG model with discriminatively learned RNNs.

This paper uses several ideas of (Socher et al., 2011b). The main differences are (i) the dual representation of nodes as discrete categories and vectors, (ii) the combination with a PCFG, and (iii) the syntactic untying of weights based on child categories. We directly compare models with fully tied and untied weights. Another work that represents phrases with a dual discrete-continuous representation is (Kartsaklis et al., 2012).

附录

并非必需，但是对于读者深入理解的的工作有帮助，往往非常形式化

- 证明
- “鸡肋”

恰当地使用附录能显著提升论文的可读性

附录

Appendix A: Table of Notation

$\mathbf{f}$	source sentence
$\mathbf{f}_1^S$	sequence of source sentences: $\mathbf{f}_1, \ldots, \mathbf{f}_s, \ldots, \mathbf{f}_S$
f	source word
J	length of $\mathbf{f}$
j	position in $\mathbf{f}$, $j = 1, 2, \ldots, J$
f_j	the j-th word in $\mathbf{f}$
f_0	empty cept on the source side

Appendix B: Using the IBM Models as Feature Functions

In this article, we use IBM Models 1–4 as feature functions by taking the logarithm of the models themselves rather than the sub-models just for simplicity. It is easy to separate each sub-model as a feature as suggested by Fraser and Marcu (2006). We distinguish

写作常见问题

- 句子过长
- 经常使用被动句式
- 结构松散、口语化
- 不定冠词和定冠词的使用
- 公式后面文字的缩进
- 引用的写法

写作常见问题 （句子过长）

research communities. To accelerate the development of Chinese language processing technology, under a grant from 863 Program, Institute of Computing Technology of Chinese Academy of Sciences took part in building Corpora Resources of 863 Program together with Institute of Automation of Chinese Academy of Sciences, Tsinghua University, Peking University, Beijing HanWang Technology Corporation, Anhui USTC iFLYTEK Corporation, Graduate School of the Chinese Academy of Sciences and Institute of Linguistics of Chinese Academy of Social Sciences.

写作常见问题 （句子过长）

research communities. To accelerate the development of Chinese language processing technology, under a grant from 863 Program, Institute of Computing Technology of Chinese Academy of Sciences took part in building Corpora Resources of 863 Program together with Institute of Automation of Chinese Academy of Sciences, Tsinghua University, Peking University, Beijing HanWang Technology Corporation, Anhui USTC iFLYTEK Corporation, Graduate School of the Chinese Academy of Sciences and Institute of Linguistics of Chinese Academy of Social Sciences.

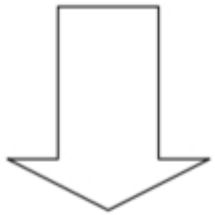
To advance the state of the art of Chinese language processing technology, many institutions in China took part in building the Corpora Resources under the grant from the 863 Program. These institutions include …

写作常见问题　（被动句式+弱动词）

The whole process of finding fuzzy-matched word pairs and computing their similarity is demonstrated in detail. More importantly, the performance of BLEU is significantly improved by integrating fuzzy matching.

写作常见问题 （被动句式+弱动词）

The whole process of finding fuzzy-matched word pairs and computing their similarity is demonstrated in detail. More importantly, the performance of BLEU is significantly improved by integrating fuzzy matching.



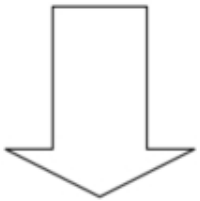
We demonstrate how to find fuzzy-matched word pairs and compute their similarities in detail. More importantly, integrating fuzzy matching significantly improved the translation performance in terms of BLEU.

写作常见问题　（结构松散+口语化）

In this step, we want to induce an alignment between words and predicates. The alignment can give a roughly mapping between words and the predicates that express their meanings, so it would be a useful constraint for rule extraction and reduce the searching space.

写作常见问题　（结构松散+口语化）

In this step, we want to induce an alignment between words and predicates. The alignment can give a roughly mapping between words and the predicates that express their meanings, so it would be a useful constraint for rule extraction and reduce the searching space.



This step induces an alignment between words and predicates. Reflecting a rough mapping between natural languages and logic, such alignments impose linguistically motivated constraints on the search space and improve the efficiency of rule extraction.

写作常见问题　（a 还是 an）

A FBI agent or *An FBI agent*?

A FIFA officer or *An FIFA officer*?

写作常见问题　（如何使用the）

The statistical translation models that try to capture the recursive structures of the language over the last several years.

In the experiments on the Chinese-English translation, we find that the model chooses to build the structures that are more syntactic.

“the”一般指特指，否则不加

写作常见问题 （如何使用the）

~~The~~ statistical translation models that try to capture the recursive structures of ~~the~~ language over the last several years.

In ~~the~~ experiments on ~~the~~ Chinese-English translation, we find that the model chooses to build ~~the~~ structures that are more syntactic.

写作常见问题 （如何使用the）

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmin}} \left\{ \sum_{s=1}^{S} E(\mathbf{r}_s, \hat{\mathbf{a}}(\mathbf{f}_s, \mathbf{e}_s; \lambda_1^M)) \right\} \tag{7}$$

$$= \underset{\lambda_1^M}{\operatorname{argmin}} \left\{ \sum_{s=1}^{S} \sum_{k=1}^{K} E(\mathbf{r}_s, \mathbf{a}_{s,k}) \delta(\hat{\mathbf{a}}(\mathbf{f}_s, \mathbf{e}_s; \lambda_1^M), \mathbf{a}_{s,k}) \right\} \tag{8}$$

where $\hat{\mathbf{a}}(\mathbf{f}_s, \mathbf{e}_s; \lambda_1^M)$ is the best candidate alignment produced by the linear model:

$$\hat{\mathbf{a}}(\mathbf{f}_s, \mathbf{e}_s; \lambda_1^M) = \underset{\mathbf{a}}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(\mathbf{f}_s, \mathbf{e}_s, \mathbf{a}) \right\} \tag{9}$$

The basic idea of MERT is to optimize only one parameter (i.e., feature weight) each time and keep all other parameters fixed. This process runs iteratively over M parameters until it cannot further reduce the loss on the training corpus.

当公式后的文本与公式有关，则不缩进，否则缩进

其它

- 论文中每个符号都应当找得到定义，除非众所周知。永远不要不加说明就是用数学符号
- 要避免数学符号冲突，使用符号列表
- 不要生造术语，尤其是中式译法，尽量符合惯例
- 集成所有信息元素，排版美观和专业

提高英语写作的窍门

- 找著名学者（尤其是native speaker）的论文钻研，学习句式和词汇用法，做笔记
- 拿不准的地方找google，双引号查询
- 学习句式和用法

提高英语写作的窍门

The need to segment and label sequences arises in many different problems in several scientific fields. Hidden

句式 *the need to … arises in … problems (fields)*

提高英语写作的窍门

The need to segment and label sequences arises in many different problems in several scientific fields. Hidden

句式 *the need to … arises in … problems (fields)*

造句

The need to learn latent-variable models from unlabeled data arises in many NLP problems.

利用搜索引擎

Maximizing the likelihood _____ the training data.

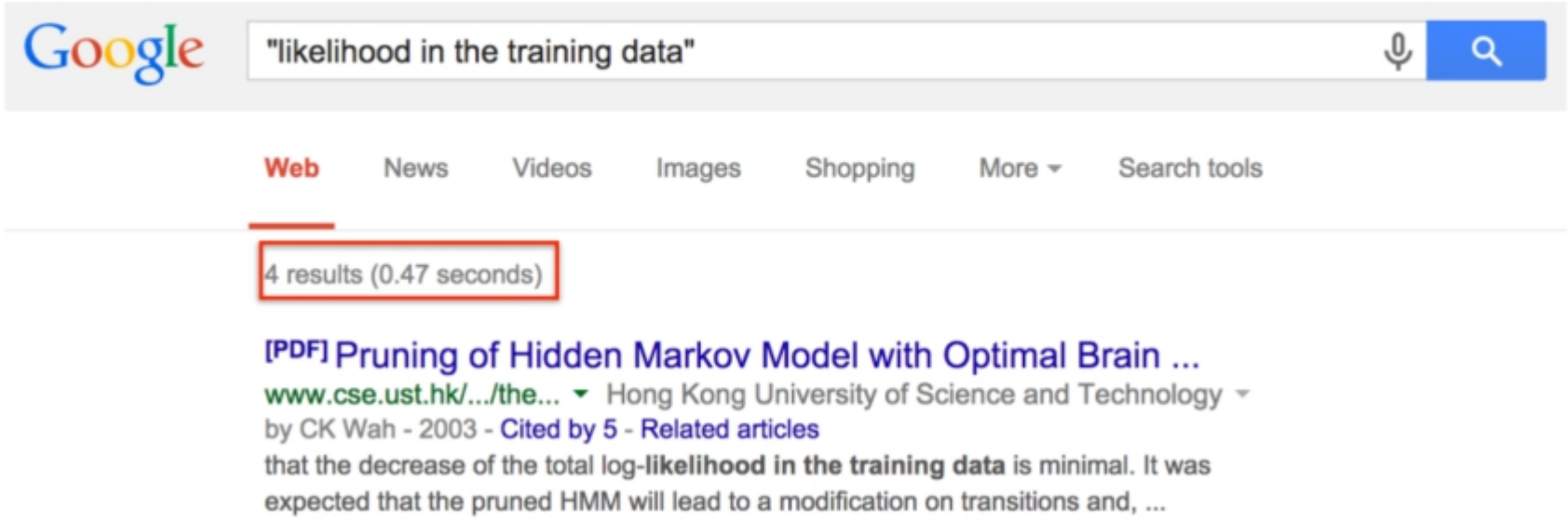
(A) in (B) on (c) of

利用搜索引擎

Maximizing the likelihood _____ the training data.

(A) in (B) on (c) of

4

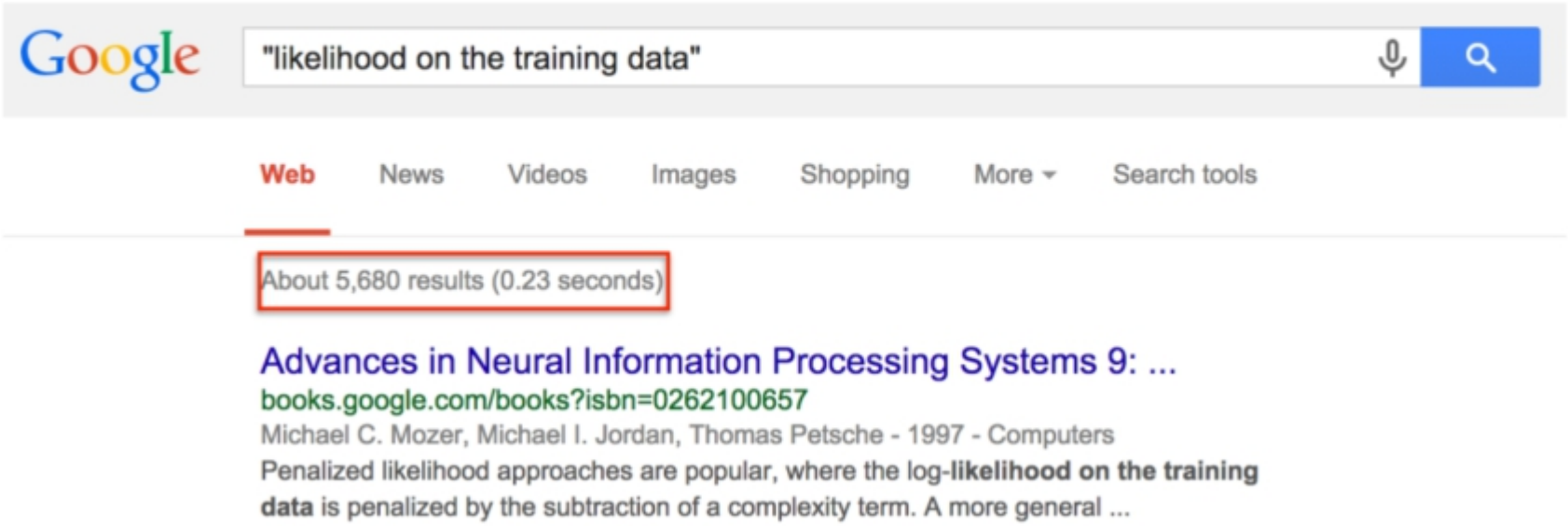


利用搜索引擎

Maximizing the likelihood _____ the training data.

(A) in (B) on (c) of

4 5,680

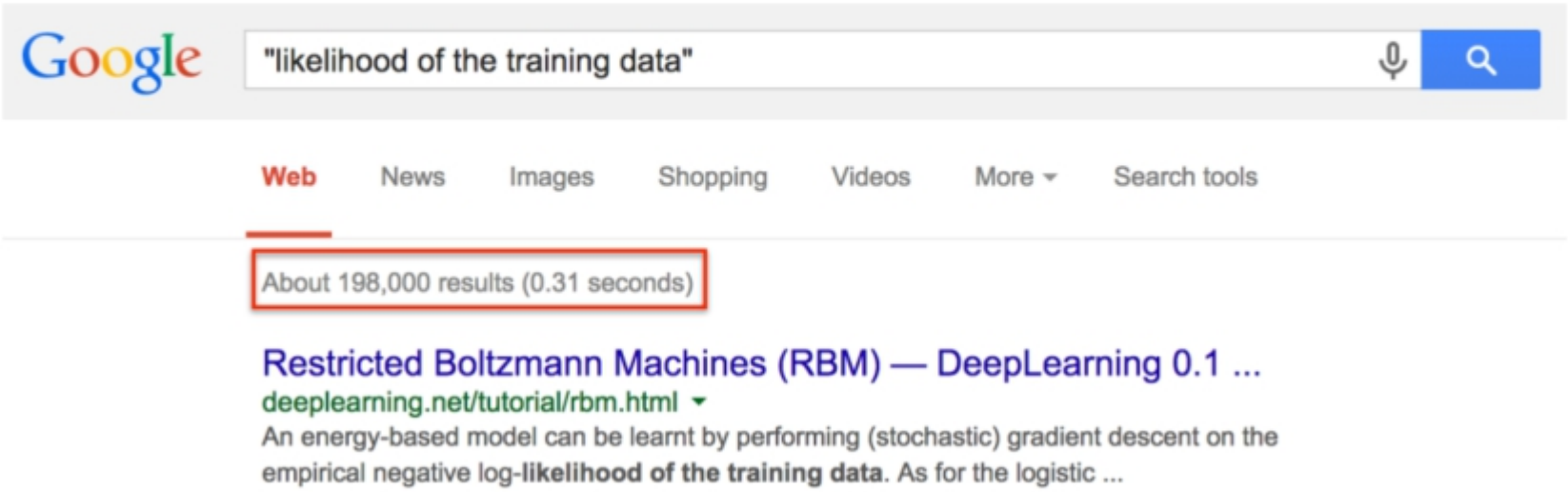


利用搜索引擎

Maximizing the likelihood _____ the training data

(A) in (B) on (c) of

4 5,680 198,000



必须掌握的工具

Latex

- 强烈建议投稿的论文使用Latex
- http://www.ctex.org/

Powerpoint

- 画矢量图

时间管理和获得反馈

Coarse-to-fine

- 截稿前一个月开始写
- 每隔两天改一次

听取不同背景读者的反馈意见

- 专家：专业意见
- 非专家：发现信息壁垒

写到极致，完成精致的艺术品

总结

论文的本质是分享思想，呈现信息

信息的呈现符合读者的认知惯性

全心全意为读者服务，降低阅读难度，提高愉悦感

细节决定成败

不要本末倒置：创新至上，技法为辅


Q&A