

清华大学学位论文 L^AT_EX 模板

使用示例文档 v5.5.2

(申请清华大学工学硕士学位论文)

培 养 单 位: 计 算 机 科 学 与 技 术 系

学 科: 计 算 机 科 学 与 技 术

研 究 生: 薛 瑞 尼

指 导 教 师: 郑 纬 民 教 授

副指导教师: 陈 文 光 教 授

联合指导教师: 某 某 某 教 授

二〇一九年六月

An Introduction to L^AT_EX Thesis Template of Tsinghua University v5.5.2

Thesis Submitted to
Tsinghua University
in partial fulfillment of the requirement
for the professional degree of
Doctor of Engineering

by
Xue Ruini
(Computer Science and Technology)

Thesis Supervisor : Professor Zheng Weimin
Associate Supervisor : Chen Wenguang

June, 2019

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：(1) 已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；(2) 为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容。

本人保证遵守上述规定。

(保密的论文在解密后应遵守此规定)

作者签名：_____

导师签名：_____

日 期：_____

日 期：_____

摘 要

论文的摘要是对论文研究内容和成果的高度概括。摘要应对论文所研究的问题及其研究目的进行描述，对研究方法和过程进行简单介绍，对研究成果和所得结论进行概括。摘要应具有独立性和自明性，其内容应包含与论文全文同等量的主要信息。使读者即使不阅读全文，通过摘要就能了解论文的总体内容和主要成果。

论文摘要的书写应力求精确、简明。切忌写成对论文书写内容进行提要的形式，尤其要避免“第 1 章……；第 2 章……；……”这种或类似的陈述方式。

本文介绍清华大学论文模板 ThuThesis 的使用方法。本模板符合学校的本科、硕士、博士论文格式要求。

本文的创新点主要有：

- 用例子来解释模板的使用方法；
- 用废话来填充无关紧要的部分；
- 一边学习摸索一边编写新代码。

关键词是为了文献标引工作、用以表示全文主要内容信息的单词或术语。关键词不超过 5 个，每个关键词中间用分号分隔。（模板作者注：关键词分隔符不用考虑，模板会自动处理。英文关键词同理。）

关键词：TeX；LaTeX；CJK；模板；论文

Abstract

An abstract of a dissertation is a summary and extraction of research work and contributions. Included in an abstract should be description of research topic and research objective, brief introduction to methodology and research process, and summarization of conclusion and contributions of the research. An abstract should be characterized by independence and clarity and carry identical information with the dissertation. It should be such that the general idea and major contributions of the dissertation are conveyed without reading the dissertation.

An abstract should be concise and to the point. It is a misunderstanding to make an abstract an outline of the dissertation and words “the first chapter”, “the second chapter” and the like should be avoided in the abstract.

Key words are terms used in a dissertation for indexing, reflecting core information of the dissertation. An abstract may contain a maximum of 5 key words, with semi-colons used in between to separate one another.

Key Words: TeX; LaTeX; CJK; template; thesis

Contents

Chapter 1 Introduction	1
1.1 Motivation	1
1.2 Outline	1
1.3 Contributions	1
1.3.1 基本表格	1
1.3.2 复杂表格	2
1.3.3 其它	4
1.4 定理环境	5
1.5 参考文献	7
1.6 公式	8
Chapter 2 Background	10
2.1 Coreference Resolution	10
2.1.1 Related Work	10
2.2 Pronoun Resolution	11
2.3 Gender Bias	11
2.3.1 插图	12
Chapter 3 Method	14
3.1 Mind the GAP	14
3.2 Extract Features with BERT	14
3.3 Fine-tune BERT	14
3.4 Evaluation Method	14
3.4.1 Multi-class Logarithmic Loss	14
3.4.2 F1 Score	14
Chapter 4 Experiments and Results	15
4.1 Data	15
4.2 Extract features with BERT	15
4.3 Fine-tune BERT	15
Chapter 5 Conclusions	16
Bibliography	17

Contents

Acknowledgments	19
声 明	20

Nomenclature

NLU	Natural Language Understanding
cluster	集群
Itanium	安腾
SMP	对称多处理
API	应用程序编程接口
PI	聚酰亚胺
MPI	聚酰亚胺模型化合物, N-苯基邻苯酰亚胺
PBI	聚苯并咪唑
MPBI	聚苯并咪唑模型化合物, N-苯基苯并咪唑
PY	聚吡咯
PMDA-BDA	均苯四酸二酐与联苯四胺合成的聚吡咯薄膜
ΔG	活化自由能 (Activation Free Energy)
χ	传输系数 (Transmission Coefficient)
E	能量
m	质量
c	光速
P	概率
T	时间
v	速度

Chapter 1 Introduction

1. 有一条巨龙抓走了公主 [1] (介绍你的问题为什么值得研究)
2. 巨龙是多么多么多么难打 (强调你的研究的重要性^①)
3. 王子提着一把金光闪闪的剑而不是破斧子烂长矛登场 (你的方法好在哪里, 别人 sui 在哪里)
4. 王子是如何打败巨龙 (你的方法简介)
5. 从此王子和公主幸福的生活在一起。(解决了问题)

1.1 Motivation

封面的例子请参看 `cover.tex`。主要符号表参看 `denotation.tex`, 附录和个人简历分别参看 `appendix01.tex` 和 `resume.tex`。里面的命令都很直观, 一看即会^②。

1.2 Outline

1.3 Contributions

The contributions of this thesis are summarized as follows:

1.3.1 基本表格

模板中关于表格的宏包有三个: `booktabs`、`array` 和 `longtabular`, 命令有一个 `\hline`。三线表可以用 `booktabs` 提供的 `\toprule`、`\midrule` 和 `\bottomrule`。它们与 `longtable` 能很好的配合使用。如果表格比较简单的话可以直接用命令 `\hline` 控制。

首先来看一个最简单的表格。表 4.1 列举了本模板主要文件及其功能。请大家注意三线表中各条线对应的命令。这个例子还展示了如何在表格中正确使用脚注。由于 \LaTeX 本身不支持在表格中使用 `\footnote`, 所以我们不得不将表格放在小页中, 而且最好将表格的宽度设置为小页的宽度, 这样脚注看起来才更美观。

^① ex

^② 你说还是看不懂? 怎么会呢?

Table 1.1 模板文件。如果表格的标题很长，那么在表格索引中就会很不美观，所以要像 `chapter` 那样在前面用中括号写一个简短的标题。这个标题会出现在索引中。

文件名	描述
<code>thuthesis.ins</code>	L ^A T _E X 安装文件，DocStrip ^①
<code>thuthesis.dtx</code>	所有的一切都在这里 ^② 。
<code>thuthesis.cls</code>	模板类文件。
<code>thuthesis-numeric.bst</code>	参考文献 BibT _E X 样式文件。
<code>thuthesis-author-year.bst</code>	参考文献 BibT _E X 样式文件。
<code>thuthesis.sty</code>	常用的包和命令写在这里，减轻主文件的负担。

① 表格中的脚注

② 再来一个

1.3.2 复杂表格

我们经常会在表格下方标注数据来源，或者对表格里面的条目进行解释。前面的脚注是一种不错的方法，如果不喜欢脚注，可以在表格后面写注释，比如表 1.2。

Table 1.2 复杂表格示例 1。这个引用^[2]不会导致编号混乱。

x \ y	First Half		Second Half	
	1st Qtr	2nd Qtr	3rd Qtr	4th Qtr
East*	20.4	27.4	90	20.4
West**	30.6	38.6	34.6	31.6

注：数据来源《ThuThesis 使用手册》。

*：东部

**：西部

此外，表 1.2 同时还演示了另外两个功能：1) 通过 `tabularx` 的 `|X|` 扩展实现表格自动放大；2) 通过命令 `\diagbox` 在表头部分插入反斜线。

浮动体的并排放置一般有两种情况：1) 二者没有关系，为两个独立的浮动体；2) 二者隶属于同一个浮动体。对表格来说并排表格既可以像图 1.3、图 1.4 使用小页环境，也可以如图 1.5 使用子表格来做。图的例子参见第 2.3.1.2 节。

Table 1.3 第一个并排子表格

111	222
222	333

Table 1.4 第二个并排子表格

111	222
222	333

Table 1.5 并排子表格

(a) 第一个子表格		(b) 第二个子表格	
111	222	111	222
222	333	222	333

不可否认 \LaTeX 的表格功能没有想象中的那么强大，不过只要足够认真，足够细致，同样可以排出来非常复杂非常漂亮的表格。请参看表 1.6。

Table 1.6 复杂表格示例 2

Network Topology		# of nodes	# of clients			Server
GT-ITM	Waxman Transit-Stub	600	2%	10%	50%	Max. Connectivity
Inet-2.1		6000				
Xue	Rui	Ni	ThuThesis			
	ABCDEF					

Table 1.7 实验数据

测试程序	正常运行 时间 (s)	同步 时间 (s)	检查点 时间 (s)	卷回恢复 时间 (s)	进程迁移 时间 (s)	检查点 文件 (KB)
CG.A.2	23.05	0.002	0.116	0.035	0.589	32491
CG.A.4	15.06	0.003	0.067	0.021	0.351	18211
CG.A.8	13.38	0.004	0.072	0.023	0.210	9890
CG.B.2	867.45	0.002	0.864	0.232	3.256	228562
CG.B.4	501.61	0.003	0.438	0.136	2.075	123862
CG.B.8	384.65	0.004	0.457	0.108	1.235	63777
MG.A.2	112.27	0.002	0.846	0.237	3.930	236473
MG.A.4	59.84	0.003	0.442	0.128	2.070	123875
MG.A.8	31.38	0.003	0.476	0.114	1.041	60627
MG.B.2	526.28	0.002	0.821	0.238	4.176	236635
MG.B.4	280.11	0.003	0.432	0.130	1.706	123793
MG.B.8	148.29	0.003	0.442	0.116	0.893	60600
LU.A.2	2116.54	0.002	0.110	0.030	0.532	28754
LU.A.4	1102.50	0.002	0.069	0.017	0.255	14915
LU.A.8	574.47	0.003	0.067	0.016	0.192	8655
LU.B.2	9712.87	0.002	0.357	0.104	1.734	101975
LU.B.4	4757.80	0.003	0.190	0.056	0.808	53522
LU.B.8	2444.05	0.004	0.222	0.057	0.548	30134
EP.A.2	123.81	0.002	0.010	0.003	0.074	1834
EP.A.4	61.92	0.003	0.011	0.004	0.073	1743

续下页

续表 1.7 实验数据

测试程序	正常运行 时间 (s)	同步 时间 (s)	检查点 时间 (s)	卷回恢复 时间 (s)	进程迁移 时间 (s)	检查点 文件 (KB)
EP.A.8	31.06	0.004	0.017	0.005	0.073	1661
EP.B.2	495.49	0.001	0.009	0.003	0.196	2011
EP.B.4	247.69	0.002	0.012	0.004	0.122	1663
EP.B.8	126.74	0.003	0.017	0.005	0.083	1656

要想用好论文模板还是得提前学习一些 $\text{T}_\text{E}\text{X}/\text{L}_\text{A}\text{T}_\text{E}\text{X}$ 的相关知识，具备一些基本能力，掌握一些常见技巧，否则一旦遇到问题还真是比较麻烦。我们见过很多这样的同学，一直以来都是使用 Word 等字处理工具，以为 $\text{L}_\text{A}\text{T}_\text{E}\text{X}$ 模板的用法也应该类似，所以就沿袭同样的思路来对待这种所见非所得的排版工具，结果被折腾的焦头烂额，疲惫不堪。

如果您要排版的表格长度超过一页，那么推荐使用 `longtable` 或者 `supertabular` 宏包，模板对 `longtable` 进行了相应的设置，所以用起来可能简单一些。表 1.7 就是 `longtable` 的示例。

1.3.3 其它

如果不想让某个表格或者图片出现在索引里面，请使用命令 `\caption*`。这个命令不会给表格编号，也就是出来的只有标题文字而没有“表 XX”，“图 XX”，否则索引里面序号不连续就显得不伦不类，这也是 $\text{L}_\text{A}\text{T}_\text{E}\text{X}$ 里星号命令默认的规则。

有这种需求的多是本科同学的英文资料翻译部分，如果觉得附录中英文原文中的表格和图片显示成“表”和“图”不协调的话，一个很好的办法就是用 `\caption*`，参数随便自己写，比如不守规矩的表 1.111 和图 1.111 能满足这种特殊需要（可以参看附录部分）。

表 1.111 这是一个手动编号，不出现在索引中的表格。

ThuThesis

Figure 1.111 这是一个手动编号，不出现在索引中的图。

薛瑞尼

如果的确想让它编号，但又不想让它出现在索引中的话，目前模板上不支持。

最后，虽然大家不一定会独立使用小页，但是关于小页中的脚注还是有必要提一下。请看下面的例子。

柳宗元，字子厚（773-819），河东（今永济县）人^①，是唐代杰出的文学家，哲学家，同时也是一位政治改革家。与韩愈共同倡导唐代古文运动，并称韩柳^②。

① 山西永济水饺。

② 唐宋八大家之首二位。

1.4 定理环境

给大家演示一下各种和证明有关的环境：

Assumption 1.1: 待月西厢下，迎风户半开；隔墙花影动，疑是玉人来。

$$c = a^2 - b^2 \quad (1-1)$$

$$= (a + b)(a - b) \quad (1-2)$$

千辛万苦，历尽艰难，得有今日。然相从数千里，未曾衰戚。今将渡江，方图百年欢笑，如何反起悲伤？（引自《杜十娘怒沉百宝箱》）

Definition 1.1: 子曰：「道千乘之国，敬事而信，节用而爱人，使民以时。」

Proposition 1.1: 曾子曰：「吾日三省吾身——为人谋而不忠乎？与朋友交而不信乎？传不习乎？」

Remark 1.1: 天不言自高，水不言自流。

$$\begin{aligned} \varphi(x, z) &= z - \gamma_{10}x - \gamma_{mn}x^m z^n \\ &= z - Mr^{-1}x - Mr^{-(m+n)}x^m z^n \end{aligned}$$

$$\zeta^0 = (\xi^0)^2, \quad (1-3)$$

$$\zeta^1 = \xi^0 \xi^1, \quad (1-4)$$

$$\zeta^2 = (\xi^1)^2, \quad (1-5)$$

Axiom 1.1: 两点间直线段距离最短。

$$x \equiv y + 1 \pmod{m^2} \quad (1-6)$$

$$x \equiv y + 1 \pmod{m^2} \quad (1-7)$$

$$x \equiv y + 1 \pmod{m^2} \quad (1-8)$$

Lemma 1.1: 《猫和老鼠》是我最爱看的动画片。

$$\begin{aligned} & \int_a^b \left\{ \int_a^b [f(x)^2 g(y)^2 + f(y)^2 g(x)^2] - 2f(x)g(x)f(y)g(y) dx \right\} dy \\ &= \int_a^b \left\{ g(y)^2 \int_a^b f^2 + f(y)^2 \int_a^b g^2 - 2f(y)g(y) \int_a^b fg \right\} dy \end{aligned}$$

行行重行行，与君生别离。相去万余里，各在天一涯。道路阻且长，会面安可知。胡马依北风，越鸟巢南枝。相去日已远，衣带日已缓。浮云蔽白日，游子不顾返。思君令人老，岁月忽已晚。弃捐勿复道，努力加餐饭。

Theorem 1.1: 犯我强汉者，虽远必诛

—— 陈汤（汉）

$$y = 1 \quad (1-9a)$$

$$y = 0 \quad (1-9b)$$

道可道，非常道。名可名，非常名。无名天地之始；有名万物之母。故常无，欲以观其妙；常有，欲以观其徼。此两者，同出而异名，同谓之玄。玄之又玄，众妙之门。上善若水。水善利万物而不争，处众人之所恶，故几于道。曲则全，枉则直，洼则盈，敝则新，少则多，多则惑。人法地，地法天，天法道，道法自然。知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者寿。

proof 燕赵古称多感慨悲歌之士。董生举进士，连不得志于有司，怀抱利器，郁郁适兹土，吾知其必有合也。董生勉乎哉？

夫以子之不遇时，苟慕义强仁者，皆爱惜焉，矧燕、赵之士出乎其性者哉！然吾尝闻风俗与化移易，吾恶知其今不异于古所云邪？聊以吾子之行卜之也。董生勉乎哉？

吾因子有所感矣。为我吊望诸君之墓，而观于其市，复有昔时屠狗者乎？为我谢曰：“明天子在上，可以出而仕矣！”

—— 韩愈 《送董邵南序》 □

Corollary 1.1: 四川话配音的《猫和老鼠》是世界上最好看最好听最有趣的动画片。

$$V_i = v_i - q_i v_j, \quad X_i = x_i - q_i x_j, \quad U_i = u_i, \quad \text{for } i \neq j; \quad (1-10)$$

$$V_j = v_j, \quad X_j = x_j, \quad U_j u_j + \sum_{i \neq j} q_i u_i. \quad (1-11)$$

迢迢牵牛星，皎皎河汉女。纤纤擢素手，札札弄机杼。终日不成章，泣涕零如雨。河汉清且浅，相去复几许。盈盈一水间，脉脉不得语。

Example 1.1: 大家来看这个例子。

$$\begin{cases} \nabla f(\mathbf{x}^*) - \sum_{j=1}^p \lambda_j \nabla g_j(\mathbf{x}^*) = 0 \\ \lambda_j g_j(\mathbf{x}^*) = 0, \quad j = 1, 2, \dots, p \\ \lambda_j \geq 0, \quad j = 1, 2, \dots, p. \end{cases} \quad (1-12)$$

Exercise 1.1: 请列出 Andrew S. Tanenbaum 和 W. Richard Stevens 的所有著作。

Conjecture 1.1: *Poincare Conjecture* If in a closed three-dimensional space, any closed curves can shrink to a point continuously, this space can be deformed to a sphere.

Problem 1.1: 回答还是不回答，是个问题。

如何引用定理 1.1 呢? 加上 `\label` 使用 `\ref` 即可。妾发初覆额，折花门前剧。郎骑竹马来，绕床弄青梅。同居长干里，两小无嫌猜。十四为君妇，羞颜未尝开。低头向暗壁，千唤不一回。十五始展眉，愿同尘与灰。常存抱柱信，岂上望夫台。十六君远行，瞿塘滟滪堆。五月不可触，猿声天上哀。门前迟行迹，一一生绿苔。苔深不能扫，落叶秋风早。八月蝴蝶来，双飞西园草。感此伤妾心，坐愁红颜老。

1.5 参考文献

当然参考文献可以直接写 `\bibitem`，虽然费点功夫，但是好控制，各种格式可以自己随意改写。

本模板推荐使用 `BIBTEX`，分别提供数字引用 (`thuthesis-numeric.bst`) 和作者年份引用 (`thuthesis-author-year.bst`) 样式，基本符合学校的参考文献格式 (如专利等引用未加详细测试)。看看这个例子，关于书的^[2-4]，还有这些^[5-7]，关于杂志的^[8-10]，硕士论文^[11-12]，博士论文^[13-14]，标准文件^[15]，会议论文^[16-17]，技术报告^[18]，电子文献^[19-20]。若使用著者-出版年制，中文参考文献^[21] 应增加 `key={pinyin}` 字段，以便正确进行排序^[22]。另外，如果对参考文献有不如意的地方，请手动修改 `bbl` 文件。

有时候不想要上标，那么可以这样 [13]，这个非常重要。

有时候一些参考文献没有纸质出处，需要标注 URL。缺省情况下，URL 不会在连字符处断行，这可能使得用连字符代替空格的网址分行很难看。如果需要，可以将模板类文件中

`\RequirePackage{hyperref}`

一行改为:

`\PassOptionsToPackage{hyphens}{url}`

`\RequirePackage{hyperref}`

使得连字符处可以断行。更多设置可以参考 `url` 宏包文档。

1.6 公式

贝叶斯公式如式 (1-13)，其中 $p(y|\mathbf{x})$ 为后验； $p(\mathbf{x})$ 为先验；分母 $p(\mathbf{x})$ 为归一化因子。

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}, y)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})} \quad (1-13)$$

论文里面公式越多， \TeX 就越 happy。再看一个 `amsmath` 的例子：

$$\det \mathbf{K}(t = 1, t_1, \dots, t_n) = \sum_{I \in n} (-1)^{|I|} \prod_{i \in I} t_i \prod_{j \in I} (D_j + \lambda_j t_j) \det \mathbf{A}^{(\lambda)}(\bar{I}|\bar{I}) = 0. \quad (1-14)$$

前面定理示例部分列举了很多公式环境，可以说把常见的情况都覆盖了，大家在写公式的时候一定要好好看 `amsmath` 的文档，并参考模板中的用法：

$$\begin{aligned} & \int_a^b \left\{ \int_a^b [f(x)^2 g(y)^2 + f(y)^2 g(x)^2] - 2f(x)g(x)f(y)g(y) dx \right\} dy \\ &= \int_a^b \left\{ g(y)^2 \int_a^b f^2 + f(y)^2 \int_a^b g^2 - 2f(y)g(y) \int_a^b fg \right\} dy \end{aligned}$$

其实还可以看看这个多级规划：

$$\left\{ \begin{array}{l} \max_{\mathbf{x}} F(\mathbf{x}, y_1^*, y_2^*, \dots, y_m^*) \\ \text{subject to:} \\ G(\mathbf{x}) \leq 0 \\ (y_1^*, y_2^*, \dots, y_m^*) \text{ solves problems } (i = 1, 2, \dots, m) \\ \left\{ \begin{array}{l} \max_{\mathbf{x}} f_i(\mathbf{x}, y_1, y_2, \dots, y_m) \\ \text{subject to:} \\ g_i(\mathbf{x}, y_1, y_2, \dots, y_m) \leq 0. \end{array} \right. \end{array} \right. \quad (1-15)$$

这些跟规划相关的公式都来自于刘宝碇老师《不确定规划》的课件。

Chapter 2 Background

2.1 Coreference Resolution

Coreference resolution is the task typically framed as identifying all mentions of entities and events in text and clustering them into equivalence classes [23]. It is a task of identifying clusters of mentions referring to the same real-world entity. Furthermore, our machines even need to understand the coreference between sentences. It is an important step for a lot of higher level NLP tasks that involve natural language understanding such as document summarization, question answering, and information extraction.

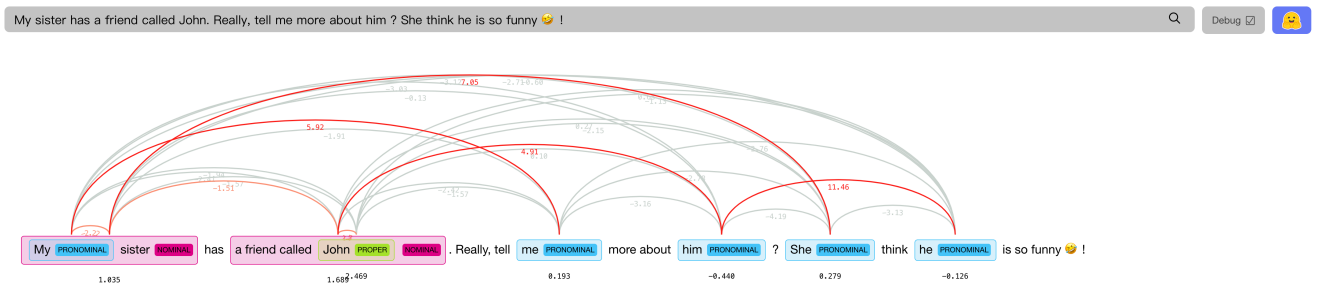


Figure 2.1 Coreference resolution visualized with HuggingFace demo <https://huggingface.co/coref/>

2.1.1 Related Work

Rule-based In the absence of large-scale data for training coreference models, early systems relied heavily on expert knowledge. A frequently used example of this is the Stanford multi-pass sieve system (Lee et al., 2011). A deterministic system, the sieve consists of multiple rule-based models which are applied in succession, from highest-precision to lowest. Gender is among the set of mention attributes identified in the very first stage of the sieve, making this information available throughout the system.

Statistical Statistical methods, often with millions of parameters, ultimately surpassed the performance of rule-based systems on shared task data (Durrett and Klein, 2013; Bjorkelund and Kuhn, 2014). The system of Durrett and Klein (2013) replaced hand-written rules with simple feature templates. Combinations of these features implicitly capture linguistic phenomena useful for resolving antecedents, but they may also unintentionally capture bias in the data. For instance, for occupations which are not frequently found in the data, an occupation+pronoun feature can be highly informative, and

the overly confident model can exhibit strong bias when applied to a new domain.

Neural The move to deep neural models led to more powerful antecedent scoring functions, and the subsequent learned feature combinations resulted in new state-of-the-art performance (Wiseman et al., 2015; Clark and Manning, 2016b). Global inference over these models further improved performance (Wiseman et al., 2016; Clark and Manning, 2016a), but from the perspective of potential bias, the information available to the model is largely the same as in the statistical models. A notable exception is in the case of systems which make use of pre-trained word embeddings (Clark and Manning, 2016b), which have been shown to contain bias and have the potential to introduce bias into the system.

2.2 Pronoun Resolution

Pronoun resolution is part of coreference resolution, the task of identifying for a specified pronoun in a passage, which named entity antecedent it refers to. This is an important task for natural language understanding and a necessary component of machine translation systems, chat bots and assistants.

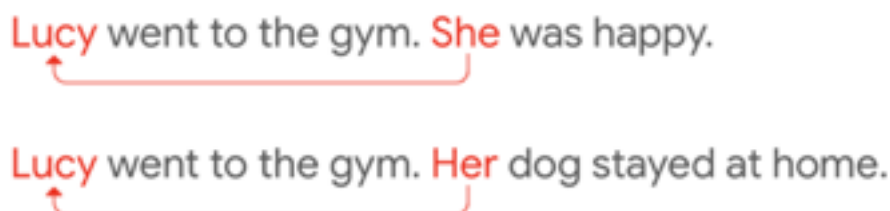


Figure 2.2 Pronoun resolution

2.3 Gender Bias

在第 1 章中我们学习了贝叶斯公式 (1-13)，这里我们复习一下：

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}, y)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})} \quad (2-1)$$

2.3.1 插图

强烈推荐《 $\text{\LaTeX} 2_{\epsilon}$ 插图指南》! 关于子图形的使用细节请参看 `subcaption` 宏包的说明文档。

2.3.1.1 一个图形

一般图形都是处在浮动环境中。之所以称为浮动是指最终排版效果图形的位置不一定与源文件中的位置对应^①，这也是刚使用 \LaTeX 同学可能遇到的问题。如果要强制固定浮动图形的位置，请使用 `float` 宏包，它提供了 `[H]` 参数，比如图 ??。



Figure 2.3 Pronoun Resolution

2.3.1.2 多个图形

如果多个图形相互独立，并不共用一个图形计数器，那么用 `minipage` 或者 `parbox` 就可以。否则，请参看图 2.4，它包含两个小图，分别是图 2.4(a)和图 2.4(b)。推荐使用 `\subcaptionbox`，因为可以像图 2.4 那样对齐子图的标题，也可以使用 `subcaption` 宏包的 `\subcaption` (放在 `minipage` 中,用法同`\caption`)或是 `subfigure`、`subtable` 环境，像图 2.5，不要再用 `\subfloat`、`\subfigure` 和 `\subtable`。



(a) 第一个小图形



(b) 第二个小图形，注意这个图略矮些。如果标题很长的话，它会自动换行

Figure 2.4 包含子图形的大图形 (`subcaptionbox` 示例)



(a) 第一个小图形



(b) 第二个小图形，注意这个图略矮些。`subfigure` 中同一行的子图在顶端对齐。

Figure 2.5 包含子图形的大图形 (`subfigure` 示例)

^① This is not a bug, but a feature of \LaTeX !



Figure 2.6 并排第一个图



Figure 2.7 并排第二个图

如果要把编号的两个图形并排，那么小页就非常有用：

李氏子蟠，年十七，好古文、六艺，经传皆通习之，不拘於时，学於余。余嘉其能行古道，作师说以贻之。

—— 韩愈（唐）

Chapter 3 Method

3.1 Mind the GAP

3.2 Extract Features with BERT

3.3 Fine-tune BERT

3.4 Evaluation Method

3.4.1 Multi-class Logarithmic Loss

The evaluation metric for the following project is the multi-class logarithmic loss as shown below.

$$\text{logloss} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{ij} \log(p_{ij}) \quad (3-1)$$

where N is the number of samples in the test set, M is 3, \log is the natural logarithm, y_{ij} is 1 if observation i belongs to class j and 0 otherwise, and p_{ij} is the predicted probability that observation i belongs to class j .

The probabilities are not required to sum to one because they are rescaled prior to being scored (each row is divided by the row sum). In order to avoid the extremes of the log function, predicted probabilities are replaced with $\max(\min(p, 1 - 10^{-15}), 10^{-15})$.

3.4.2 F1 Score

In order to measure gender bias, in addition to an overall F_1 score O , we also use separate F_1 scores for the masculine, M , and feminine, F , examples and take a feminine-to-masculine F_1 score ratio, F/M , as a measure of bias as done in the original GAP paper [24].

Chapter 4 Experiments and Results

4.1 Data

Table 4.1 GAP Dataset Overview

Column	Header	Description
1	ID	Unique identifier for an example (two pairs)
2	Text	Text containing the ambiguous pronoun and two candidate names. About a paragraph in length
3	Pronoun	The pronoun, text
4	Pronoun-offset	Character offset of Pronoun in Column 2 (Text)
5	A	The first name, text
6	A-offset	Character offset of A in Column 2 (Text)
7	A-coref	Whether A corefers with the pronoun, TRUE or FALSE
8	B	The second name, text
9	B-offset	Character offset of B in Column 2 (Text)
10	B-coref	Whether B corefers with the pronoun, TRUE or FALSE
11	URL	The URL of the source Wikipedia page

Following Kaggle competition “Gendered Pronoun Resolution”, 4 for each abstract from Wikipedia pages we are given a pronoun, and we try to predict the right coreference for it, i.e. to which named entity (A or B) it refers. Let’s take a look at this simple example: “John entered the room and saw [A] Julia. [Pronoun] She was talking to [B] Mary Hendriks and looked so extremely gorgeous that John was stunned and couldn’t say a word.” Here “Julia” is marked as entity A, “Mary Hendriks” – as entity B, and pronoun “She” is marked as Pronoun. In this particular case the task is to correctly identify to which entity the given pronoun refers.

4.2 Extract features with BERT

4.3 Fine-tune BERT

Chapter 5 Conclusions

Bibliography

- [1] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
- [2] Knuth D E. The \TeX book[M]. 15th ed. Reading, MA: Addison-Wesley Publishing Company, 1989.
- [3] Goosens M, Mittelbach F, Samarin A. The \LaTeX companion[M]. Reading, MA: Addison-Wesley Publishing Company, 1994: 112-125.
- [4] Gröning P, Nilsson L, Ruffieux P, et al. Encyclopedia of nanoscience and nanotechnology: volume 1[M]. USA: American Scientific Publishers, 2004: 547-579.
- [5] Krasnogor N. Towards robust memetic algorithms[M]//Hart W, Krasnogor N, Smith J. Studies in Fuzziness and Soft Computing: volume 166 Recent Advances in Memetic Algorithms. New York: Springer Berlin Heidelberg, 2004: 185-207.
- [6] 阎真. 沧浪之水[M]. 北京: 人民文学出版社, 2001: 185-207.
- [7] 班固. 苏武传[M]//郑在瀛, 汪超宏, 周文复. 新古文观止丛书: 第2卷 传记散文英华. 武汉: 湖北人民出版社, 1998: 65-69.
- [8] Chafik El Idrissi M, Roney A, Frigon C, et al. Measurements of total kinetic-energy released to the $N = 2$ dissociation limit of H_2 — evidence of the dissociation of very high vibrational Rydberg states of H_2 by doubly-excited states[J]. Chemical Physics Letters, 1994, 224(10):260-266.
- [9] Mellinger A, Vidal C R, Jungen C. Laser reduced fluorescence study of the carbon-monoxide and triplet Rydberg series-experimental results and multichannel quantum-defect analysis[J]. J. Chem. Phys., 1996, 104(5):8913-8921.
- [10] Shell M. How to use the IEEEtran \LaTeX class[J]. Journal of \LaTeX Class Files, 2002, 12(4): 100-120.
- [11] 猪八戒. 论流体食物的持久保存[硕士学位论文]. 北京: 广寒宫大学, 2005.
- [12] Jeyakumar A R. Metamori: A library for incremental file checkpointing[D]. Blacksburg: Virginia Tech, 2004.
- [13] 沙和尚. 论流沙河的综合治理[博士学位论文]. 北京: 清华大学, 2005.
- [14] Zadok E. FiST: A System for Stackable File System Code Generation[D]. USA: Computer Science Department, Columbia University, 2001.
- [15] IEEE Std 1363-2000. IEEE standard specifications for public-key cryptography[M]. New York: IEEE, 2000.
- [16] Kim S, Woo N, Yeom H Y, et al. Design and Implementation of Dynamic Process Management for Grid-enabled MPICH[C]//Dongarra J, Laforenza D, Orlando S. the 10th European PVM/MPI Users' Group Conference. Venice, Italy: Springer-Verlag, 2003.
- [17] Kocher C, Jaffe J, Jun B. Differential power analysis[C]//Wiener M. Lecture Notes in Computer Science: volume 1666 Advances in Cryptology (CRYPTO '99). Germany: Springer-Verlag, 1999: 388-397.

- [18] Woo A, Bailey D, Yarrow M, et al. The NAS parallel benchmarks 2.0[R/OL]. The Pennsylvania State University CiteSeer Archives, 1995. <http://www.nasa.org/>.
- [19] 萧钰. 出版业信息化迈入快车道[EB/OL]. (2001-12-19)[2002-04-15]. <http://www.creader.com/news/200112190019.htm>.
- [20] Online Computer Library Center, Inc. History of OCLC[EB/OL]. 2000[2000-01-08]. <http://www.oclc.org/about/history/default.htm>.
- [21] 贾宝玉, 林黛玉, 薛宝钗, 等. 论刘姥姥食量大如牛之现实意义[J]. 红楼梦杂谈, 1800, 224: 260-266.
- [22] 王重阳, 黄药师, 欧阳峰, 等. 武林高手从入门到精通[C]//第 N 次华山论剑. 西安, 中国: 金大庸, 2006.
- [23] Pradhan S, Moschitti A, Xue N, et al. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes[C]//Joint Conference on EMNLP and CoNLL-Shared Task. [S.l.]: Association for Computational Linguistics, 2012: 1-40.
- [24] Webster K, Recasens M, Axelrod V, et al. Mind the gap: A balanced corpus of gendered ambiguous pronouns[J]. Transactions of the Association for Computational Linguistics, 2018, 6: 605-617.

Acknowledgments

First and foremost, I would like to thank my supervisor, professor Miyamoto Yoichiro who always encouraged me to work hard and provided an excellent environment for research with his support and guidance.

I would also like to thank my M.Sc. supervisory committee members Dr. Wendy Os- born and Dr. Jackie Rice for spending time reading my thesis and providing valuable feedback.

I thank my beloved family who encouraged me throughout my studies at the Sophia University. Without their love and support, I would not have been here and could not finish this work.

声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名：_____ 日 期：_____