论文写作的易读性原则

案例分析：基于Seq2Seq的对话数据增广

报告人：刘一佳

合作者：侯宇泰、车万翔、刘挺

任务型对话系统中的数据增广

- 目标：帮助用户完成特定任务
- 重要子模块：语言理解 (NLU)

show me the closest restaurant.

show me the [closest]<distance> [restaurant]<poi type>

挑战：丰富的表达方式 vs 稀缺的标注数据

- 实际数据
  - show me the closest restaurant.
  - give me the closest route to a restaurant.
  - is there a close restaurant?
- 标注数据
  - show me the [closest]<distance> [restaurant]<poi type>.

挑战：丰富的表达方式 vs 稀缺的标注数据

- 实际数据
  - show me the closest restaurant.
  - give me the closest route to a restaurant.
  - is there a close restaurant?
- 标注数据
  - show me the [closest]<distance> [restaurant]<poi type>.

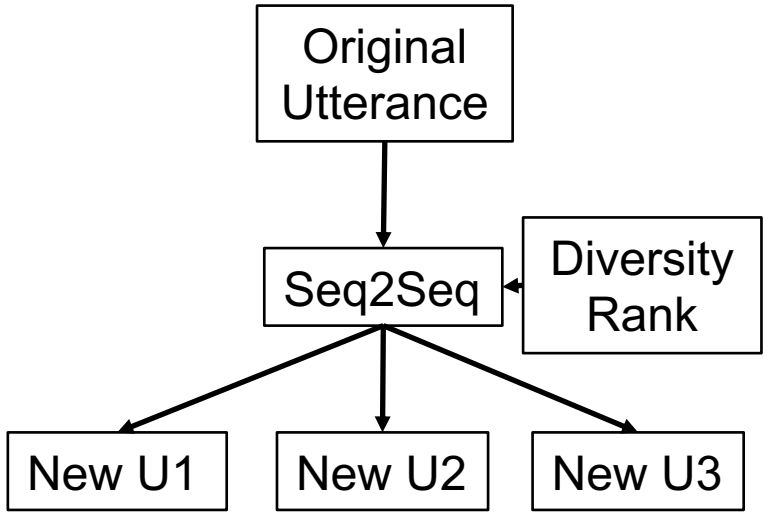
解决方案：基于Seq2Seq的数据增广

基本思想：具有相同intent和slot type的训练数据互为翻译

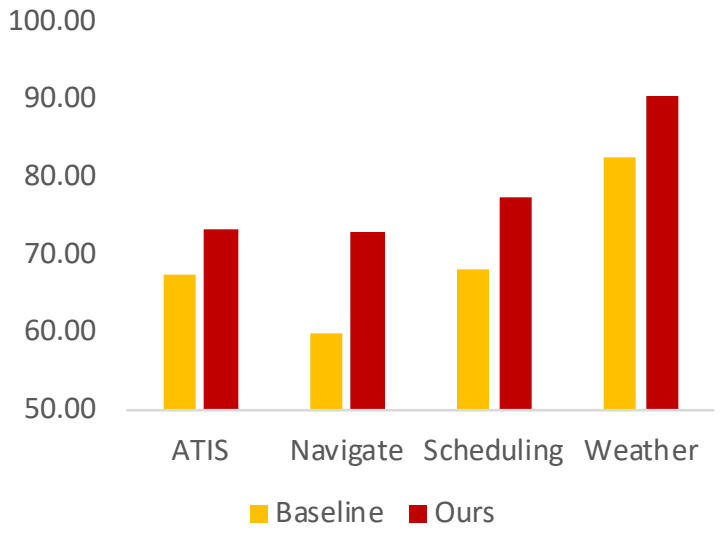
find me the <distance> route to <poi type>
is there a <distance> <poi type>
I'm desiring to eat at some <poi type> is there any in <distance>

find me the <distance> route to <poi type>
→ is there a <distance> <poi type>
find me the <distance> route to <poi type>
→ I'm desiring to eat at some <poi type> is there any in <distance>

任务型对话系统中的数据增广
总结 / 发表



方法



结果

Dear Yutai Hou:

On behalf of the COLING 2018 Program Committee, we are delighted to inform you that the following submission has been accepted to appear at the conference:

Sequence-to-Sequence Data Augmentation for Dialogue Language Understanding

COLING score: 4/4/4

发表

HIT-SCIR

摘要

第一版

Task-oriented dialogue systems built by dialogue system development platform, like api.ai and Microsoft LUIS, usually suffer from lack of both user utterance and data annotations at the early stage of development, which results in poor performance facing the real user. To solve this, we propose Cluster2Cluster, a generation model that augments the data by creating new data instances of cluster level. Since user utterance fed to dialogue systems is naturally clustered by intent and data slots, we leverage this information to direct generation by first clustering user utterance and then expanding utterance clusters with a generation process that directly uses the clusters as input and output. Our model outperforms strong baseline of bi-LSTM with existing data augmentation method and gets satisfactory results of generating diverse new utterance. To the best of our knowledge, we are first to leverage clustering information in data augmentation for natural language processing and first to propose cluster level natural language generation model.

最终版

In this paper, we study the problem of data augmentation for language understanding in task-oriented dialogue system. In contrast to previous work which augments an utterance without considering its relation with other utterances, we propose a sequence-to-sequence generation based data augmentation framework that leverages one utterance's same semantic alternatives in the training data. A novel diversity rank is incorporated into the utterance representation to make the model produce diverse utterances and these diversely augmented utterances help to improve the language understanding module. Experimental results on the Airline Travel Information System dataset and a newly created semantic frame annotation on Stanford Multi-turn, Multi-domain Dialogue Dataset show that our framework achieves significant improvements of 6.38 and 10.04 F-scores respectively when only a training set of hundreds utterances is represented. Case studies also confirm that our method generates diverse utterances.

第一版

Task-oriented dialogue system [*talk about task-oriented dialogue*] With the increasing demand for task oriented dialogue [*talk about commercial systems like api.ai*]. However, in the common scenario of building new dialogue system for unfamiliar domains, developer always face the lack of data reflected both in quantity and quality. [*talk about quantity and quality*]

To solve this, we leverage the idea of data augmentation. In CV and speech, [*literature review, CV methods don't work for NLP*] Therefore, existing data augmentation methods for NLP generally follow two directions [*more literature review*] However, the prevailing methods do not well introduce domain knowledge of specific tasks, and their effectiveness tends to be limited in the respective tasks.

To remedy this, more information should be used to direct data augmentation. [*we noticed they can be naturally clustered*]But, controllable generation of new data with clustering information, is [*difficult*]. There are mainly 2 challenges

---

最终版

Language understanding (LU) is the initial and essential component in the task-oriented dialogue system pipeline. One challenge in building robust LU is to handle myriad ways in which users express demands. [*talk about new domain cases*]

Data augmentation, which enlarges the size of training data in machine learning systems, is an effective solution to the data insufficiency problem. [*talk about successful cases in CV, speech, and NLP*] but less studied in LU. Kurata et al. (2016a) present [*a data driven, non-paraphrasing method, contrast it to paraphrasing method, pointed they need domain knowledge*]

In this paper, we study the problem of data augmentation for LU and propose a novel data-driven framework that models relations between utterances of the same semantic frame in the training data. [*talk about our method*]

论文写作的易读性原则


Bla Bla Bla Bla Bla Bla Bla Bla Bla Bla Bla Bla Bla Bla Bla Bla Bla Bla

论文写作易读性原则

- 宏观一致性
- 微观一致性
- 奥卡姆剃刀
- 建模读者
- 符合学术共识

论文写作易读性原则

- 宏观一致性：文章是围绕同一个主题组织的
- 微观一致性
- 奥卡姆剃刀
- 建模读者
- 符合学术共识

原则：宏观一致性

第一版

Task-oriented dialogue system [*talk about task-oriented dialogue*] With the increasing demand for task oriented dialogue [*talk about commercial systems like api.ai*]. However, in the common scenario of building new dialogue system for unfamiliar domains, developer always face the lack of data reflected both in quantity and quality. [*talk about quantity and quality*]

To solve this, we leverage the idea of data augmentation. In CV and speech, [*literature review, CV methods don't work for NLP*] Therefore, existing data augmentation methods for NLP generally follow two directions [*more literature review*] However, the prevailing methods do not well introduce domain knowledge of specific tasks, and their effectiveness tends to be limited in the respective tasks.

To remedy this, more information should be used to direct data augmentation. [*we noticed they can be naturally clustered*] But, controllable generation of new data with clustering information, is [*difficult*]. There are mainly 2 challenges

---

最终版

Language understanding (LU) is the initial and essential component in the task-oriented dialogue system pipeline. One challenge in building robust LU is to handle myriad ways in which users express demands. [*talk about new domain cases*]

Data augmentation, which enlarges the size of training data in machine learning systems, is an effective solution to the data insufficiency problem. [*talk about successful cases in CV, speech, and NLP*] but less studied in LU. Kurata et al. (2016a) present [*a data driven, non-paraphrasing method, pointed it doesn't make use of relations between utterances*]

In this paper, we study the problem of data augmentation for LU and propose a novel data-driven framework that models relations between utterances of the same semantic frame in the training data. [*talk about our method*]

原则：宏观一致性

- 错误的示范
  - “我们研究了对话系统中的数据库检索问题
  - “Memory Network近年取得了比较广泛的应用，[100字描述它如何广泛应用]，我们用它建模数据库检索。”

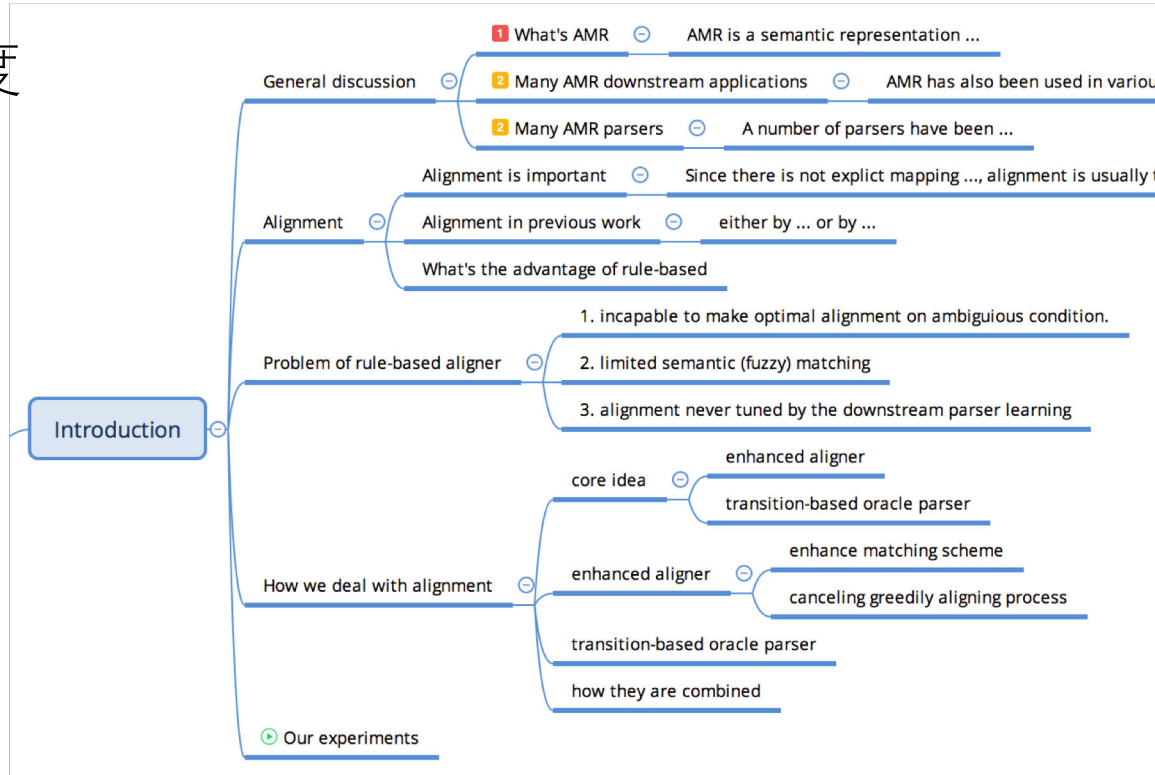
- 比较正确的示范
  - “我们研究了对话系统中的数据库检索问题
  - “检索问题可以建模为A问题。Memory Network近年在A问题取得了比较广泛的应用，我们用它建模数据库检索。”

原则：宏观一致性

- 宏观一致性的表现：文章是围绕同一个主题组织的

- 提高宏观一致性的方法：**重复**
  - 重复不等于复制：一个主题，多种角度

- 提高宏观一致性的工具：**思维导图**
  - “我的文章的主题是什么？”
  - “我的这段与主题是什么关系？”



论文写作易读性原则

- 宏观一致性
- 微观一致性：细节
- 奥卡姆剃刀
- 建模读者
- 符合学术共识

原则：微观一致性

Training Dialogues	2,424
Validation Dialogues	302
Test Dialogues	304
Navigation Dialogues	1000
Scheduling Dialogues	1034
Weather Dialogues	996
Avg. # of Utterances Per Dialogue	5.25
Avg.	9
Vocabulary Size	1601

格式不一致

不一致会带来关注

原则：微观一致性

- 微观一致性的表现
  - 句子主从句主语保持一致
  - 段落各句主语尽量一致
  - 使用相同的专有名词，符号指代相同的事物
  - 小数点位数，表格的居中方式，图片位置

- 提高微观一致性的方法
  - 多轮修改
  - 找其他人proofreading
  - 形成习惯

论文写作易读性原则

- 宏观一致性
- 微观一致性
- 奥卡姆剃刀：文章是用最简单的方式写作的
- 建模读者
- 符合学术共识

原则：奥卡姆剃刀

- 错误的示范

$$p(C_{out}|C_{in}, CM) = \sum_{s \in C_{out}} p(s \mid C_{in}, CM)$$

$$\approx \sum_{s \in C_{out}} \sum_{s' \in C_{in}} p(s \mid s', CM)$$

引入了一个无意义的记号

$$p(x|y) = \prod_{t=1}^{l} p(y_t \mid X, y^1, y^2, ..., y^{t-1})$$

原则：奥卡姆剃刀

- 错误的示范
  - “对一个句子做数据增广”
    - 把句子聚成一类C
      - [这里又有建模读者的问题，读到这里读者会问和哪些句子聚类、怎么聚类]
    - 从类C里预测一个集合的句子D

- 比较正确的示范
  - “对一个句子做数据增广”
    - 这个句子输入到A模型得到一个集合的句子B

原则：奥卡姆剃刀

- 提高简洁性的方法
  - “这个长句能否变成短句”
  - “这个方法有没有更简单的表述形式？”

- 提高简洁性的训练
  - 用一句话概括一篇文章
  - 用一分钟给别人讲懂一篇文章

论文写作易读性原则

- 宏观一致性
- 微观一致性
- 奥卡姆剃刀
- 建模读者：写作时考虑读者的理解到的内容
- 符合学术共识

原则：建模读者

Task-oriented dialogue systems built by dialogue system development platform, like api.ai and Microsoft LUIS, usually suffer from lack of both user utterance and data annotations at the early stage of development, which results in poor performance facing the real user. To solve this, we propose Cluster2Cluster, a generation model that augments the data by creating new data instances of cluster level. Since user utterance fed to dialogue systems is naturally clustered by intent and data slots, we leverage this information to direct generation by first clustering user utterance and then expanding utterance clusters with a generation process that directly uses the clusters as input and output. Our model outperforms strong baseline of bi-LSTM with existing data augmentation method and gets satisfactory results of generating diverse new utterance. To the best of our knowledge, we are first to leverage clustering information in data augmentation for natural language processing and first to propose cluster level natural language generation model.

原则：建模读者

Task-oriented dialogue systems built by dialogue system development platform, like api.ai and Microsoft LUIS, usually suffer from lack of both user utterance and data annotations at the early stage of development, which results in poor performance facing the real user. To solve this, we propose Cluster2Cluster, a generation model that augments the data by creating new data instances of cluster level. Since user utterance fed to dialogue systems is naturally clustered by intent and data slots, we leverage this information to direct generation by first clustering user utterance and then expanding utterance clusters with a generation process that directly uses the clusters as input and output. Our model outperforms strong baseline of bi-LSTM with existing data augmentation method and gets satisfactory results of generating diverse new utterance. To the best of our knowledge, we are first to leverage clustering information in data augmentation for natural language processing and first to propose cluster level natural language generation model.

- Task oriented dialogue
- Commercial systems

原则：建模读者

Task-oriented dialogue systems built by dialogue system development platform, like api.ai and Microsoft LUIS, usually suffer from lack of both user utterance and data annotations at the early stage of development, which results in poor performance facing the real user. To solve this, we propose Cluster2Cluster, a generation model that augments the data by creating new data instances of cluster level. Since user utterance fed to dialogue systems is naturally clustered by intent and data slots, we leverage this information to direct generation by first clustering user utterance and then expanding utterance clusters with a generation process that directly uses the clusters as input and output. Our model outperforms strong baseline of bi-LSTM with existing data augmentation method and gets satisfactory results of generating diverse new utterance. To the best of our knowledge, we are first to leverage clustering information in data augmentation for natural language processing and first to propose cluster level natural language generation model.

- Task oriented dialogue
- Commercial systems
- Data insufficiency
- Generation at cluster-level
  - What cluster?

原则：建模读者

Task-oriented dialogue systems built by dialogue system development platform, like api.ai and Microsoft LUIS, usually suffer from lack of both user utterance and data annotations at the early stage of development, which results in poor performance facing the real user. To solve this, we propose Cluster2Cluster, a generation model that augments the data by creating new data instances of cluster level. Since user utterance fed to dialogue systems is naturally clustered by intent and data slots, we leverage this information to direct generation by first clustering user utterance and then expanding utterance clusters with a generation process that directly uses the clusters as input and output. Our model outperforms strong baseline of bi-LSTM with existing data augmentation method and gets satisfactory results of generating diverse new utterance. To the best of our knowledge, we are first to leverage clustering information in data augmentation for natural language processing and first to propose cluster level natural language generation model.

- Task oriented dialogue
- Commercial systems
- Data insufficiency
- Generation at cluster-level
  - What cluster?
- Intent and slot
  - Where are they from?

原则：建模读者

Task-oriented dialogue systems built by dialogue system development platform, like api.ai and Microsoft LUIS, usually suffer from lack of both user utterance and data annotations at the early stage of development, which results in poor performance facing the real user. To solve this, we propose Cluster2Cluster, a generation model that augments the data by creating new data instances of cluster level. Since user utterance fed to dialogue systems is naturally clustered by intent and data slots, we leverage this information to direct generation by first clustering user utterance and then expanding utterance clusters with a generation process that directly uses the clusters as input and output. Our model outperforms strong baseline of bi-LSTM with existing data augmentation method and gets satisfactory results of generating diverse new utterance. To the best of our knowledge, we are first to leverage clustering information in data augmentation for natural language processing and first to propose cluster level natural language generation model.

- Task oriented dialogue
- Commercial systems
- Data insufficiency
- Generation at cluster-level
  - What cluster?
- Intent and slot
  - Where are they from?
- Cluster utterances
  - How?

原则：建模读者

- Task oriented dialogue
- Commercial systems
- Data insufficiency
- Generation at cluster-level
  - What cluster?
- Intent and slot
  - Where are they from?
- Cluster utterances
  - How?

向读者传递了一个无用的概念

读者还没有建立起cluster的概念

Intent和slot是标注数据 但前文并没有建立起增广标注数据的概念

原则：建模读者

- 建模读者的表现
  - “读者读到这里，有没有建立起A概念”
  - “读者看到Fig B的引用，能否很快地找到Fig B”
  - “读者读到这里，会不会有疑问C，我是不是需要解释”
  - “我最希望我的读者关注D部分，现在读到D了，他高不高兴”

- 提高建模读者能力的方法
  - draft需要更多人反馈
  - rebuttal积累经验

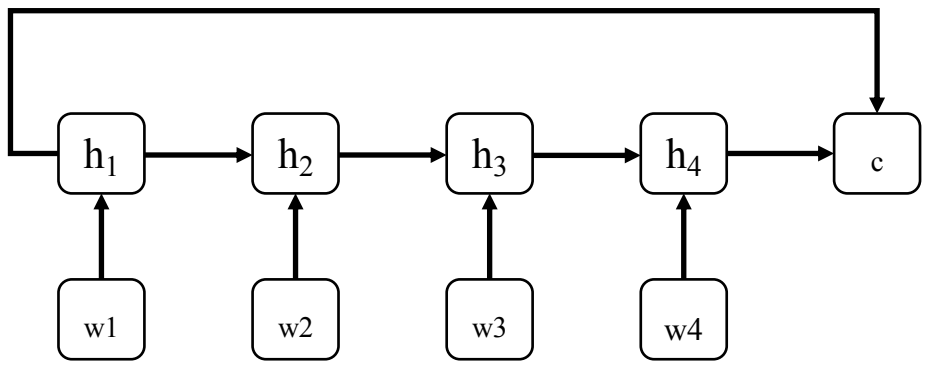
论文写作易读性原则

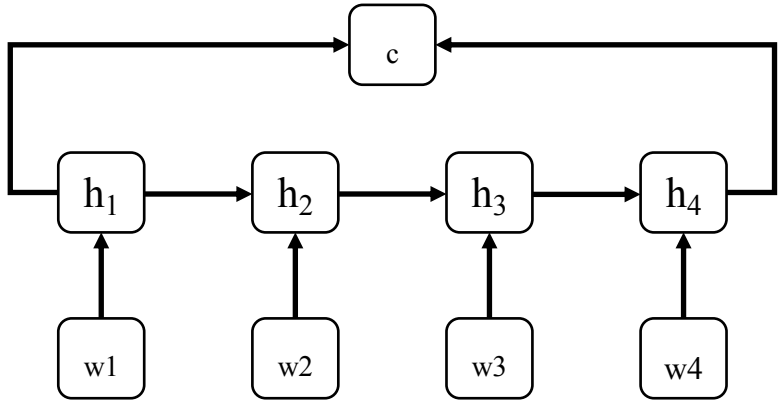
- 宏观一致性
- 微观一致性
- 奥卡姆剃刀
- 建模读者
- 符合学术共识

原则：符合学术共识

- 错误的示范
  - “我们将情感分析建模为从句子到类别标识的生成问题”

- 正确的示范
  - “我们将情感分析建模为分类问题”





原则：符合学术共识

Cluster-to-cluster?

我们把一个训练集(cluster)变成另一个训练集(cluster)了

但这个变换是不是分解成一系列独立的句子级问题了？

是

那和seq2seq有什么区别？

好像没区别

那你觉得我们在cluster-to-cluster上有什么贡献吗？

或者cluster-to-cluster应该是什么样的模型？

我觉得如果叫cluster-to-cluster，cluster内的句子的生成应该是不独立的

原则：符合学术共识

- 积累学术共识的方法
  - 刷论文
  - 交流

论文写作易读性原则

原则	表现	提高方法
宏观一致性	文章是围绕同一个主题组织的	重复、思维导图
微观一致性	细节	多轮修改
奥卡姆剃刀	文章是用最简单的方式写作的	概括能力的训练
建模读者	写作时考虑读者的理解到的内容	找人修改、多投稿
符合学术共识		刷论文、交流

论文写作必读

第十届全国机器翻译研讨会

中国澳门 2014年11月

机器翻译学术论文写作方法和技巧

刘洋

论文写作之外：人的因素

- “找其他人proofreading”
- “建模读者”
- “利用学术共识”

- 保持健康的人际关系是产出优秀工作的重要因素

论文写作之外：健康的身体

祝大家产出优秀的学术工作