基于变分自编码器的半监督任务型语言理解

Semi-supervised Training Using Variational Auto-encoder For Spoken Language Understanding

答辩人：赵知非

指导教授：何炎祥 彭敏

目录

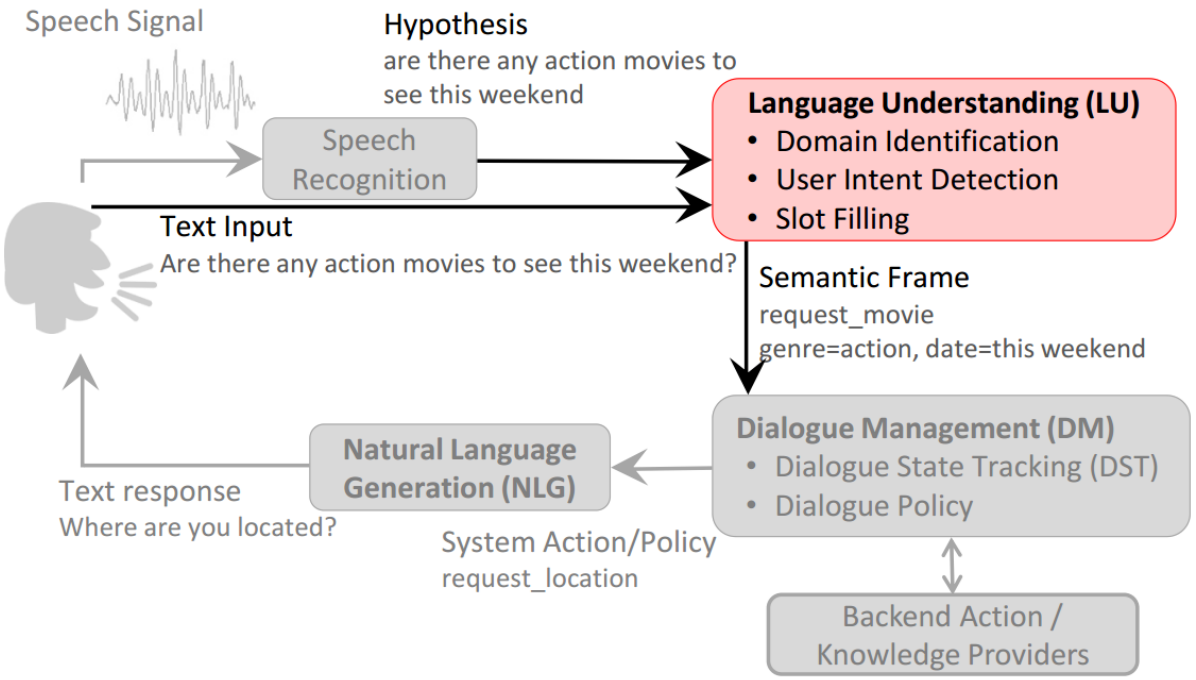
1 研究背景 Research Background

小爱同学 Hey Siri

Google ASSISTANT Cortana

语言理解 (Language Understanding，LU/SLU/NLU)

在任务型对话系统中，语言理解技术能将用户输入的自然语言解析为与特定任务相关的**语义表示**，该语义表示包含了最能代表说话者意图的语义单元。



Query： “播放周杰伦的七里香”

1.领域识别(Domain Identification)

"music" 领域

2.意图检测(Intent Detection)

"play_music"意图

3.语义槽填充(Semantic Slot Filling)

“播放[O] / 周杰伦[B-singer] / 的[O] / 七里香[B-song]"

1 研究背景 Research Background

领域识别与意图检测

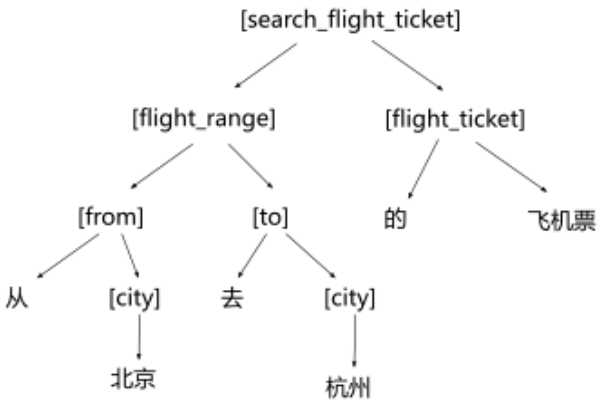
本质上是文本分类问题

Tur et al., 2012

Sarikaya et al., 2011　　Ravuri et al., 2015　　Lee et al,2016

CFG　　SVM/ME　　DBN　　DCN　　RNN　　RNN+CNN

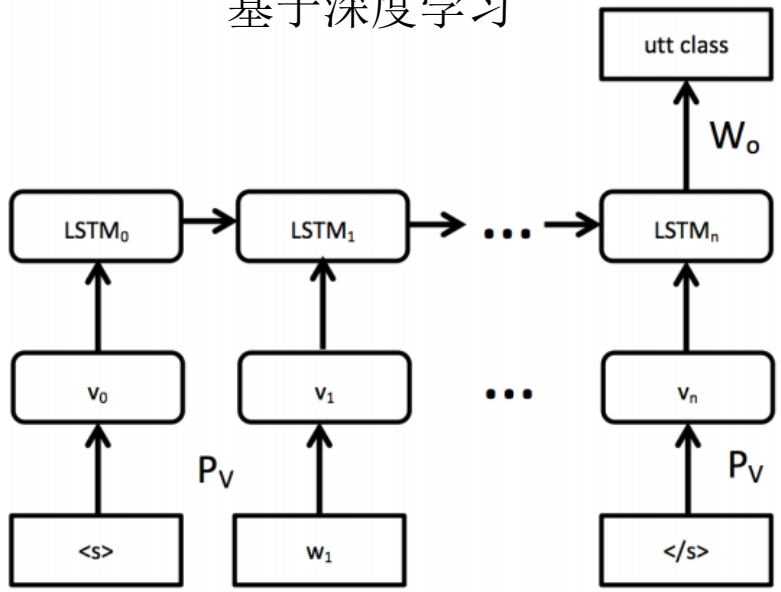
基于规则方法



基于传统机器学习

- Bag of words 提取基本特征
- 根据预定义规则去除stopwords，提取意图词等高质量特征
- 利用tf-idf将特征表示为向量
- 利用特征向量训练SVM等分类器

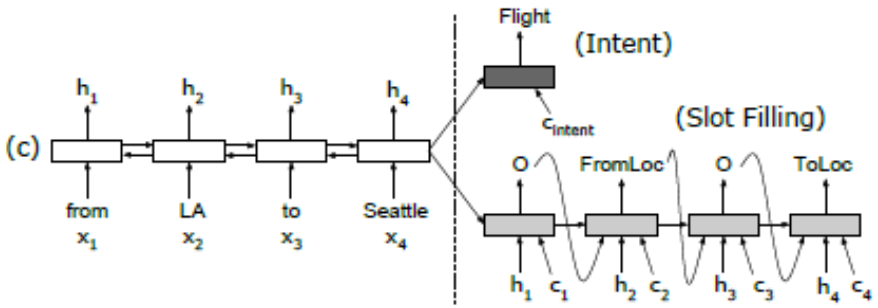
基于深度学习



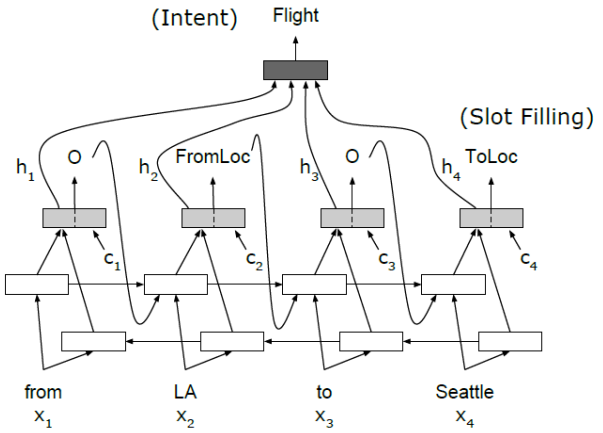
1 研究背景 Research Background

语义槽填充

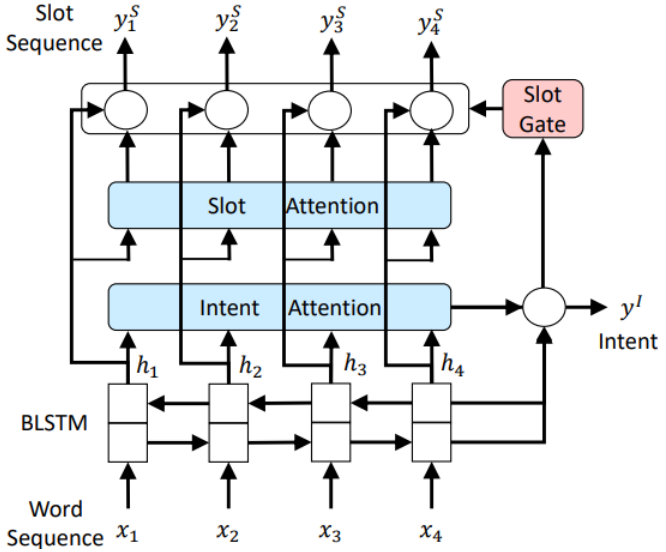
本质上是序列标注问题



Attention Encoder-Decoder（2016）



Attention BiRNN（2016）



Slot-Gated Model with Full Attention（2018）

$$c_i = \sum_{j=1}^{T} \alpha_{i,j} h_j \qquad \alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=1}^{T} \exp(e_{i,k})}$$

$$e_{i,k} = g(s_{i-1}, h_k)$$

c : 上下文向量(context vector)
h : 隐层对齐向量

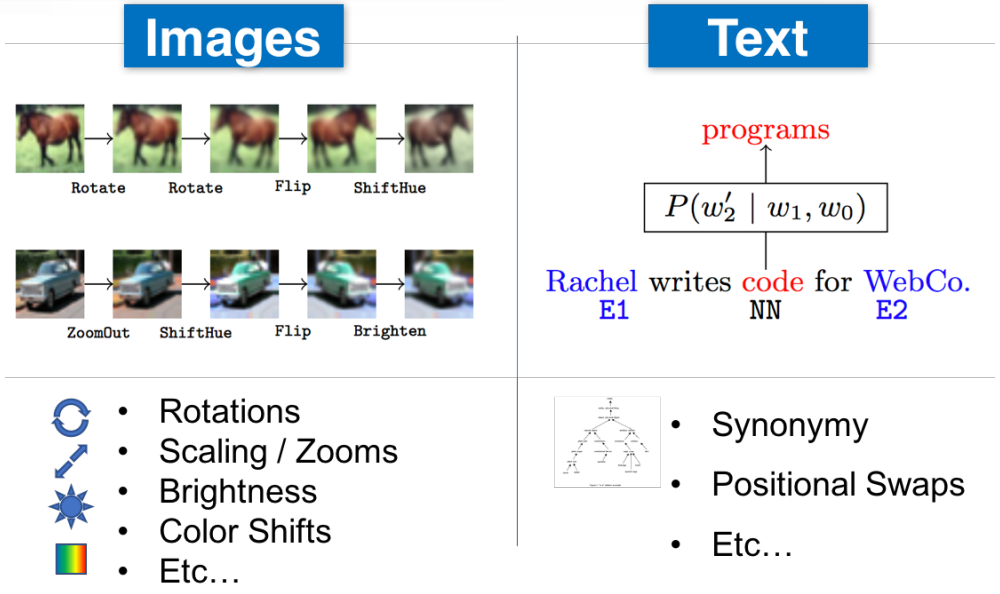
1 研究背景 Research Background

联合模型面临的数据稀缺问题

- 难以处理用户表达需求的各种表述方式
- 新兴领域难以得到大规模标注数据
- 数据收集采样过程中的偏差性
- **标注数据匮乏，标注质量难以控制**

数据问题带来的弊端

1. 扩充数据集 => 高训练成本（大量的标注时间、人力消耗）
2. 使用小数据集训练 => 模型泛化能力不强（过拟合问题）
3. 利用传统DA方法生成数据集 => 生成数据缺乏多样性与鲁棒性（完全依赖于DA函数）



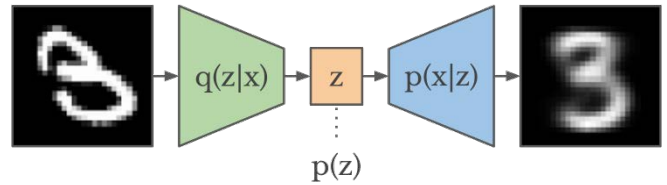
传统Data Augmentation方法

本文目的

- 基于变分自编码器（Variation Autoencoder, VAE）提出一种新的用于任务型对话系统中的半监督语言理解模型。

2 相关工作 Related Work

语言理解领域的半监督学习 *Semi-supervised Learning in SLU*



1

VAE提出

- 最早由Kingma等人在2014年提出应用于**图像领域**的半监督变分自编码器方法

2

预训练+有监督学习

- Celikyilmaz等人于2016年对未标记数据采用了**伪标记技术**来进行预处理
- Matthew等人于2017年采用了**双向语言模型**来提高序列标记任务的准确性

3

利用VAE或GAN进行特定的语言理解任务

- GANs： Yu et al. 2017（文本生成*Text generation*）； Fedus, Goodfellow, and Dai 2018（文本生成）
- VAEs： Kurata, Xiang, and Zhou 2016（语义槽填充）； Hou et al. 2018（数据增强*DA*）

2 相关工作 Related Work

拟突破难点 *Semi-supervised Learning in SLU*

1.如何在变分自编码器框架下进行语义槽填充和意图识别的联合训练？

- 除了使用每个词的潜在语义进行序列标注外，在SLU领域，还需要学习句子级别的语义表示来用于意图识别。

2.如何有效利用少量标注数据和大量无标注数据的信息改进语言理解效果？

- 目前的预训练和有监督学习两阶段方法，在有监督学习阶段仍然需要利用大量标注数据来学习有效的预测模型，并没有从本质上解决标注数据集稀缺的问题。

3 研究内容

Research Content

研究目标

主要针对问题：任务型对话系统领域标注数据不足的问题

解决方案：将基于变分自编码器的半监督学习方法引入到具体的对话系统语言理解任务中

具体研究内容

1. 使用ELMO模型将数据集里所有词训练成融合上下文语义的词向量（**词=>ELMO词向量**）
2. 将训练得到的词向量作为变分自编码器中双向LSTM编码器（encoder）的输入，生成问句的潜在语义表达，并作为输入来训练意图识别的分类器（**ELMO词向量=>句子级别语义表达 => 意图识别**）
3. 将问句的潜在语义表达作为双向LSTM的输入，生成问句中每个词的潜在语义表示，并作为输入训练语义槽填充模块（**句子级别语义表达=>词语级别语义表达 => 语义槽填充**）
4. 将问句每个词的潜在语义表示作为双向LSTM解码器（decoder）的输入，生成对应的问句序列，使用随机梯度下降法对整个模型进行端到端的训练。（**词语级别语义表达 => 生成的新数据**）

3 研究内容

Research Content

创新点

1. 将半监督学习方法引入到对话系统的语言理解任务中，缓解该领域训练资源稀缺、标注成本过高的问题
2. 提出了层次变分自编码器结构，用于进行联合语义槽填充和意图识别，解决当前基于变分自编码器的方法无法进行联合语言理解的问题
3. 探索在少量标注数据情况下获得良好语言理解效果的可能

4 实验设计

Experiment Design

实验数据

数据集选择：

- ATIS（有标注）：航空旅行信息系统（ATIS）是语言理解任务中的代表性数据集
- Snips（有标注）：snips数据集是一个开源的虚拟助手语料库。数据集包含来自各个域的用户查询，例如操纵播放列表或预订餐馆
- MIT Restaurant（MR，无标注）：此数据集采集了预订餐厅相关的口头查询语句。

数据稀缺环境模拟：

按照 Chen et al.2016 的划分方式，将数据集划分为大中小(Full/Med/Small)三种规模的数据子集。

- Small：完整数据集的1/35，127 ~ 128 个问句
- Medium：完整数据集的1/9，497 ~ 498 个问句
- Full：完整数据集，4478 个问句

4 实验设计

Experiment Design

实验对比环境

对比模型：

- 基于变分自编码器的半监督语言理解模型（本文模型）
- BiLSTM（Baseline）
- Slot-Gated SLU model (Goo et al. 2018)
- Attention Encoder-Decoder，Attention BiRNN(Liu and Lane 2016)
- Deep LSTM(Kurata et al. 2016b)

对比指标：

- 不同大小的数据子集（Small/Medium/Full）下的训练效果
- 不同数据集（ATIS/Snips/MR）下的训练效果
- 不同模型在ATIS与ATIS+（数据增强后）的训练效果

5 研究计划

Research Plan

时间安排

- 2018年10月—2018年11月

  论文选题，查阅相关文献资料，撰写开题报告。

- 2018年11月—2019年1月

  模型设计，数据收集，代码编写。

- 2019年1月—2019年2月

  进行模型实验。

- 2019年1月—2019年4月

  对设计方案进行实现和模型效果验证，论文草稿撰写。

- 2019年4月—2019年5月

  根据实验结果与系统实现情况完成论文初稿。

- 2019年5月

  论文修改、定稿，参加答辩。