C2C-GenDA: Cluster-to-Cluster Generation for Data Augmentation of Slot Filling

Yutai Hou, Sanyuan Chen, Waxiang Che, Cheng Chen, Ting Liu,
Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China

1. Introduction

• Why Data Augmentation (DA)?
  (1) Spoken Language Understanding (SLU) often suffers from insufficient quantity and diversity of training data.
  (2) Data Augmentation can remedy this by automatically enlarging training set.

• What have we done?
  (1) We propose a novel Cluster-to-Cluster generation framework for Data Augmentation (DA), named C2C-GenDA.
  (2) It enforces the training set by reconstructing existing utterances into alternative expressions while keeping semantic.
  (3) Different from previous DA works that reconstruct utterances one-by-one independently, C2C-GenDA jointly encodes multiple existing utterances of the same semantics and simultaneously decodes multiple unseen expressions.

• How is the effect?
  We improve slot filling by 7.99 (11.9%) and 5.76 (13.6%) \( F\) scores on ATIS and SNIPS, when there are only hundreds of training utterances.

2. Motivation

(1) Existing DA, such as Seq2Seq re-phrasing, often unavoidably generates duplicated sentences due to the one-by-one generation.
(2) Such deficiency can inherently be avoided by cluster-to-cluster generation.

3. Method

• Cluster2Cluster generation model
  (1) Basic Cluster2Cluster Model:
    i. Input: a cluster of utterances for a certain semantic frame.
    ii. Output: a new cluster of utterances with different expressions.
  (2) Additional mechanisms to further increase diversity:
    i. Duplication-aware Attention that attends to the existing expressions to avoid duplicated decoding.
    ii. Diverse-Oriented Regularization that penalizes internal duplication within the generated cluster.

4. Experiment

• Comparison of data augmentation methods
  - Model: Baseline, NucleSeq2Seq (Koren et al., 2016), Slot Expansion (Shin, You, and Lee 2019), Rel-Seq2Seq (Hou et al., 2018), C-VAE (Shin, You, and Lee 2019).
  - Metrics: Precision, Recall, F1.

• Case Study
  - Type and use of...
    - Replace Phrases: “we show the flights from <from_city> to <to_city>” with stop in <stop_city>.
    - Enrich Info: “I'd like information on all the flights from <from_city> to <to_city>” on <depart_date>.
    - Change Syntax: “how much is a flight from <from_city> to <to_city>?”.
    - Change Semantics: “show all airlines with flights between <from_city> and <to_city>”.

• Analysis
  - Model: ATIS, Snips.
  - Metrics: Precision, Recall, F1.
  - Ablation Test: Baseline, NucleSeq2Seq, Rel-Seq2Seq, C-VAE.
  - Diversity evaluation of utterance generation.

Please list all the details related to the above content.