Better Handling Coreference Resolution in Aspect Level Sentiment Classification by Fine-Tuning Language Models

By: Dhruv Mullick,
Dr. Bilal Ghanem,
Dr. Alona Fyshe

UNIVERSITY OF ALBERTA

Outline

- **Introduction**
- Methodology
- Experiments and Results
- Conclusions

Aspect Based Sentiment Analysis

- “ABSA”
- Predicting Aspect Terms and Sentiment Polarities
- "The service was good at the restaurant, but the food was not"
    - Aspect term = “service”
      Sentiment = “positive”
    - Aspect term = “food”
      Sentiment = “negative”

Aspect Level Sentiment Classification

- “ALSC”
- Subtype of ABSA
- Predicting Sentiment for **Given** Aspect
- ("The service was good at the restaurant, but the food was not", “service”)

  -> Sentiment = “positive”

Generative Transformer Models for ABSA

- Recent work [1, 2]
- Take review as input and generate aspects with their polarities **in one go**.
- “The service was good at the restaurant, but the food was not”
  -> "service positive <sep> food negative"

[1] https://aclanthology.org/2021.acl-short.64, [2] https://aclanthology.org/2021.acl-long.188

CR Problem Observed in Generative models

("He ate food at the restaurant, it was too spicy.", food)

Expected: negative

("He ate food at the restaurant, it was deserted.", food)

Expected: neutral

Outline

- Introduction
- **Methodology**
- Experiments and Results
- Conclusions

ALSC-CR Dataset

- Measure performance of ALSC models on reviews requiring CR ability

- **CR Cases**: Reviews requiring CR ability. [Manually annotated]
  - Aspect is antecedent of the definite pronoun.
  - CR Case e.g. ("He ate food at the restaurant, it was deserted.", "restaurant").
    - Antecedent: restaurant (aspect), Pronoun: it

- **ALSC-CR** dataset
  - Test set = CR cases only.
  - Constructed from standard MAMS [1] and Rest16 [2] datasets (restaurant reviews ABSA/ALSC datasets)

[1] https://aclanthology.org/D19-1654.pdf; [2] https://aclanthology.org/S16-1002.pdf

DPR - Definite Pronoun Resolution

- Coreference Resolution Task
- Input: "The humans were afraid of the robots because *they* were strong."
  Output: "robots"
- Objective -> what is the highlighted pronoun ("they") referring to.
- Indicator of the CR ability required for ALSC-CR.

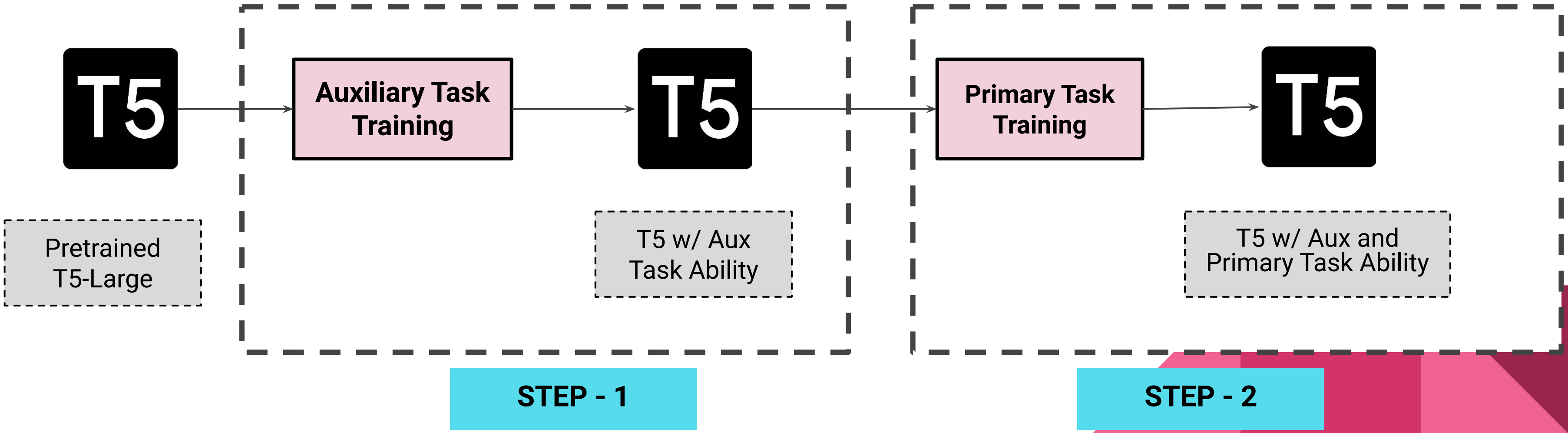
Auxiliary Tasks

- High Level Tasks
- Commonsense -> Coreference Resolution Ability

1. **CommonGen**
   Commonsense Task: Generate sentence from words
2. **CosmosQA**
   Commonsense Task: Inferential QA
3. **SQUAD**
   Extractive QA
4. **QQP**
   Quora Question Pairs. Check if two questions are semantically equivalent

Fine Tuning / Intermediate Training Setup

- 2 Step Training:
    - Fine tune with an Auxiliary task like CosmosQA, CommonGen, SQUAD, QQP
    - Fine tune with Primary task: like ALSC



Outline

- Introduction
- Methodology
- **Experiments and Results**
- Conclusions

Experiments

1. [Motivation] Show CR cases are tough by
   a. Evaluating ALSC Model (no Aux) on certain datasets
   b. Show **drop in performance for CR cases.**

2. [Solution] Show that fine tuning on Aux tasks gives improvement on ALSC-CR.

3. [Explanation] Show that Aux task improves model CR ability
   → better performance on ALSC-CR

[Motivation] CR is a Problem

- Baseline T5-large trained and evaluated on different ALSC datasets:
  - ALSC-CR
  - ALSC-Regular (not limited to CR cases in Test)

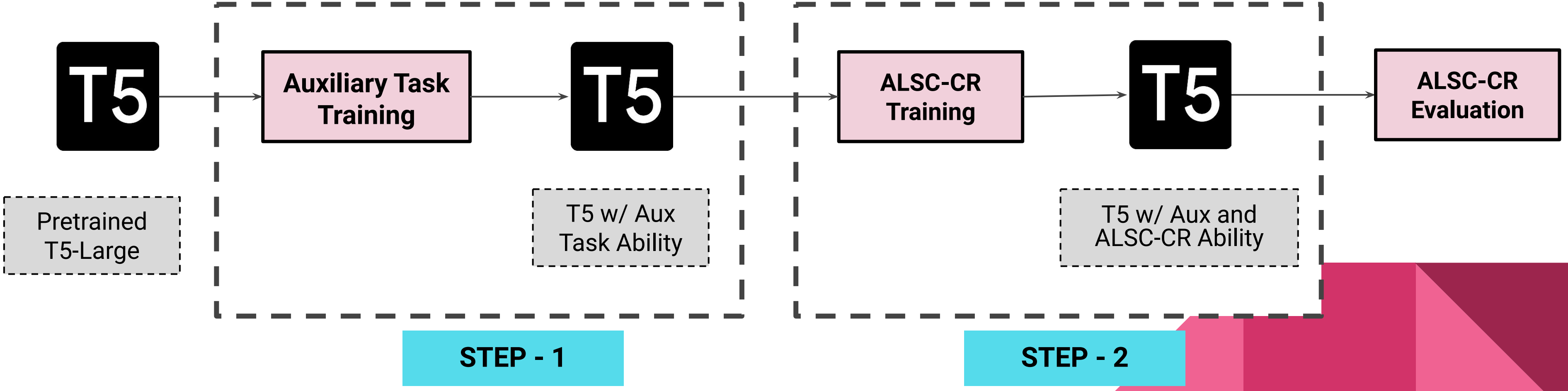
- Mean F1: Worse F1 of ALSC-CR
- Std Dev: Worse Stability of ALSC-CR

Dataset	**Mean F1 (± Std. Dev)**
ALSC-Regular	**79.71** (± 1.99)
ALSC-CR	71.07 (± 2.60)

- **Implication:** CR cases problematic for ALSC model

[Solution] Aux Tasks Improve CR Case Handling

- T5-large is fine tuned with different aux tasks before fine tuning on ALSC.
- Aux Tasks used - CommonGen, CosmosQA, SQUAD, QQP.



- QQP 0.5 and Commongen 0.1:
  - Mean F1: Improved F1
  - Std Dev: Improved Stability

- **Implication:** Using QQP 0.5 and Commongen 0.1 improves performance on ALSC-CR

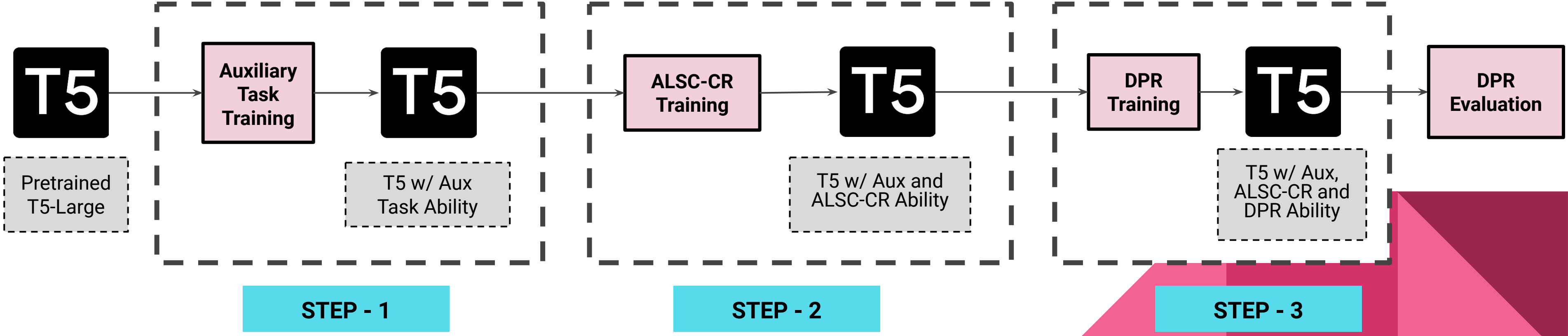
Mean-F1 (± Std Dev)

Aux. Task	**Aux. Dataset Fraction**			
	0.1	**0.2**	**0.5**	**1.0**
Commongen	75.72 (± 1.14) *	72.46 (± 2.21)	71.04 (± 3.50)	71.45 (± 1.91)
CosmosQA	71.79 (± 1.55)	71.45 (± 3.02)	72.60 (± 1.85)	73.12 (± 2.15)
SQuAD	72.02 (± 1.88)	72.60 (± 2.07)	71.47 (± 3.24)	72.08 (± 2.25)
QQP	72.49 (± 2.79)	71.85 (± 2.98)	**76.10** (± 1.26) *	71.30 (± 2.19)
N/A (Baseline)	71.07 (± 2.60)			

[Explanation] Aux Fine-tuning Improves CR Ability

- ABSA Models (Fine tuned and non Fine tuned) are trained and evaluated on DPR.
  (DPR: Identifying what highlighted pronoun refers to in given sentence)

- DPR performance of model correlates with CR ability needed for CR cases.



- QQP 0.5 and Commongen 0.1:
  - Mean F1: Improved F1
  - Std Dev: Improved Stability

Aux Task	Aux Frac.	Mean	Std Dev.
N/A (Baseline)	0	59.28	8.82
Commongen	0.1*	75.77	**1.68**
CosmosQA	1.0*	54.55	7.19
Squad	0.2	62.91	6.77
QQP	0.5*	**76.36**	2.16

- **Implication:** Using QQP 0.5 and Commongen 0.1 improves DPR (CR ability) of model.

Outline

- Introduction
- Methodology
- Experiments and Results
- **Conclusions**

Conclusions

- Handling CR Cases is a problem for generative ALSC models
- Deteriorated performance on CR Cases can be alleviated using Auxiliary fine tuning.
- Aux task fine tuning improves the CR ability which leads to performance improvement on CR Cases (ALSC-CR).

Thank You