ZOHO at SemEval-2019 Task 9: Semi-supervised Domain Adaptation using Tri-training for Suggestion Mining

Sai Prasanna, Sri Ananda Seelan
saiprasanna.r, anandaseelan.ln@zohocorp.com

Suggestion Mining

- Mining sentences that contain suggestions in online discussions and reviews.
- Example Suggestion: *"An electric kettle would have been a good addition to the room."*
- **Subtask A:** Domain specific sentence classification with training data from Microsoft Windows developer platform.
- **Subtask B**: Cross-domain classification on hotel reviews dataset.

Objective

- Evaluate recent advancements from **semi-supervised** and **transfer learning** literature to come up with a system for suggestion mining.
- Subtask A
  - Relatively small dataset
  - Class Imbalance
  - **Transfer Learning:** Use pre-trained language models and transfer it to downstream tasks.
- Subtask B
  - No hand labelled training data.
  - **Domain transfer using Semi-Supervised Learning:** Bootstrapping a model to come up with labels for data in a new domain and use it for training.

Code - https://github.com/sai-prasanna/suggestion-mining-semeval19

Built with Pytorch & AllenNLP

Tri-Training

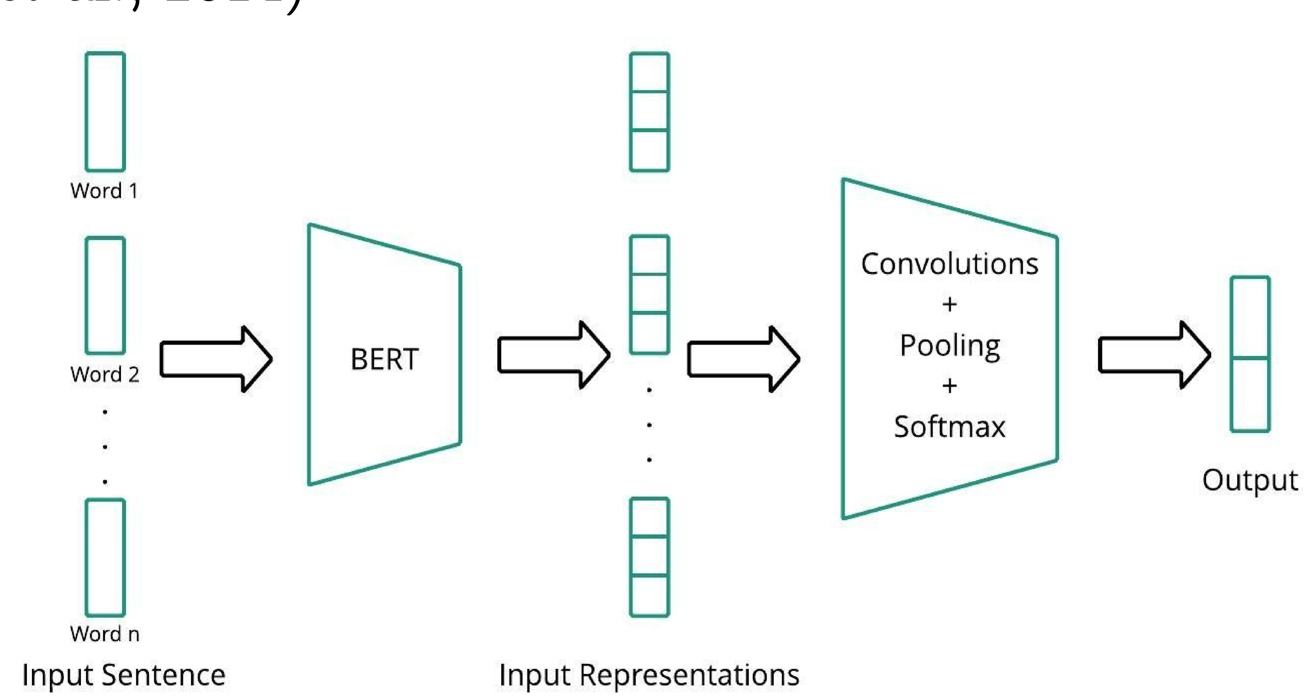
Following the work of Ruder and Plank (2018) to apply classic tri-training, a semi-supervised learning technique for domain adaptation.

Algorithm 1 Tri-training

$L \leftarrow Labelled\ Data\ ,\ |L| = m$
$U \leftarrow Unlabelled\ Data\ ,\ |U| = n$
for $i \leftarrow 1, 2, 3$ **do**
  $l_i \leftarrow BootstrapSamples(L)$
end for
repeat
  for $i \leftarrow 1, 2, 3$ **do**
    $M_i \leftarrow Train(l_i)$
  end for
  for $i \leftarrow 1, 2, 3,$ **do**
    $l_i \leftarrow L$
    for $j \leftarrow 1, n$ **do**
      if $M_p(U_j) == M_q(U_j)$
        $where\ p,\ q \neq i$ **then**
        $l_i \leftarrow l_i + \{(U_j, M_p(U_j))\}$
      end if
    end for
  end for
until no improvement in validation metrics

Model Architecture

Baseline: GloVe (Pennington et al., 2014) + Deep Averaging Net (Iyyer et al., 2015)
Final: BERT (Devlin et al., 2018) + CNN (Kim et al., 2014)



Results

Models/Experiments	Subtask - A	Subtask - B
Organizer Baseline	26.80	73.21
DAN + GloVe	38.84 ± 3.10	56.35 ± 4.71
DAN + BERT	60.82 ± 3.99	70.49 ± 4.09
CNN + BERT	64.81 ± 4.86	64.31 ± 6.72
CNN + BERT w/o Upsampling	**70.58 ± 4.24**	58.66 ± 7.79
CNN + BERT + Tritrain (Test set)	66.81 ± 1.90	**82.19 ± 1.03**
CNN + BERT + Tritrain (Yelp)	NA	81.98 ± 2.05

Table 1: F1-scores of different models/experiments
Confidence interval over 5 seeds.
DAN – Deep Averaging Network

Subtask	Model A	Model B	p-value
A	DAN + glove (Baseline)	DAN + BERT	≈ 0
A	DAN + BERT	CNN + BERT	0.046
B	CNN + BERT	CNN + BERT + Tritrain (Test set)	3.25e−08

Table 2: Pairwise comparison of various models using the **McNemar's Test**
$p \leq 0.05$ indicates a significant difference between the model performance.

References

- Ruder, S., & Plank, B. (2018). Strong Baselines for Neural Semi-Supervised Learning under Domain Shift. ACL.
- Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. ACL.
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR, abs/1810.04805.
- Pennington, J., Socher, R., & Manning, C.D. (2014). Glove: Global Vectors for Word Representation. EMNLP.
- Iyyer, M., Manjunatha, V., Boyd-Graber, J.L., & Daumé, H. (2015). Deep Unordered Composition Rivals Syntactic Methods for Text Classification. ACL.
- Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882