Seed-driven Aspect Classification on Online Reviews

Yizhe Zhao, Kening Zhang

Previous work

- Attention-based Aspect Extraction (ABAE)[1]
  - An autoencoder-based model learns a predictor by attempting to reconstruct the input sentence encoding as a linear combination of aspect embeddings.

- Teacher-student Model with Iterative Co-training[2]
  - Teacher: A Bag-of-Seed-Words Classifier
  - Student: An Embedding-Based Network
  - Refine seed word quality parameters and repeat above steps iteratively
- User-Guided Aspect Classification for Domain-Specific Texts (ARYA)[3]
  - Handle Misc aspect when tuning/expanding seed words
- Joint Aspect-Sentiment Topic Embedding (JASen)[4]
  - Learn a joint topic representation for each <sentiment, aspect> pair in the shared embedding space
  - Train word embeddings while modeling the joint distribution of seed words on all the joint topics

[1]. Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In ACL.
[2]. Giannis Karamanolakis, Daniel Hsu, and Luis Gravano.2019. Leveraging just a few keywords for finegrained aspect detection through weakly supervised co-training. In EMNLP/IJCNLP
[3]. Peiran Li, Fang Guo, and Jingbo Shang. 2020. User-Guided Aspect Classification for Domain-Specific Texts. arXiv preprint arXiv:2004.14555 (2020).
[4]. Jiaxin Huang, Yu Meng, Fang Guo, Heng Ji, and Jiawei Han. 2020. Weakly-supervised aspect-based sentiment analysis via joint aspect-sentiment topic embedding. In EMNLP.

Dataset

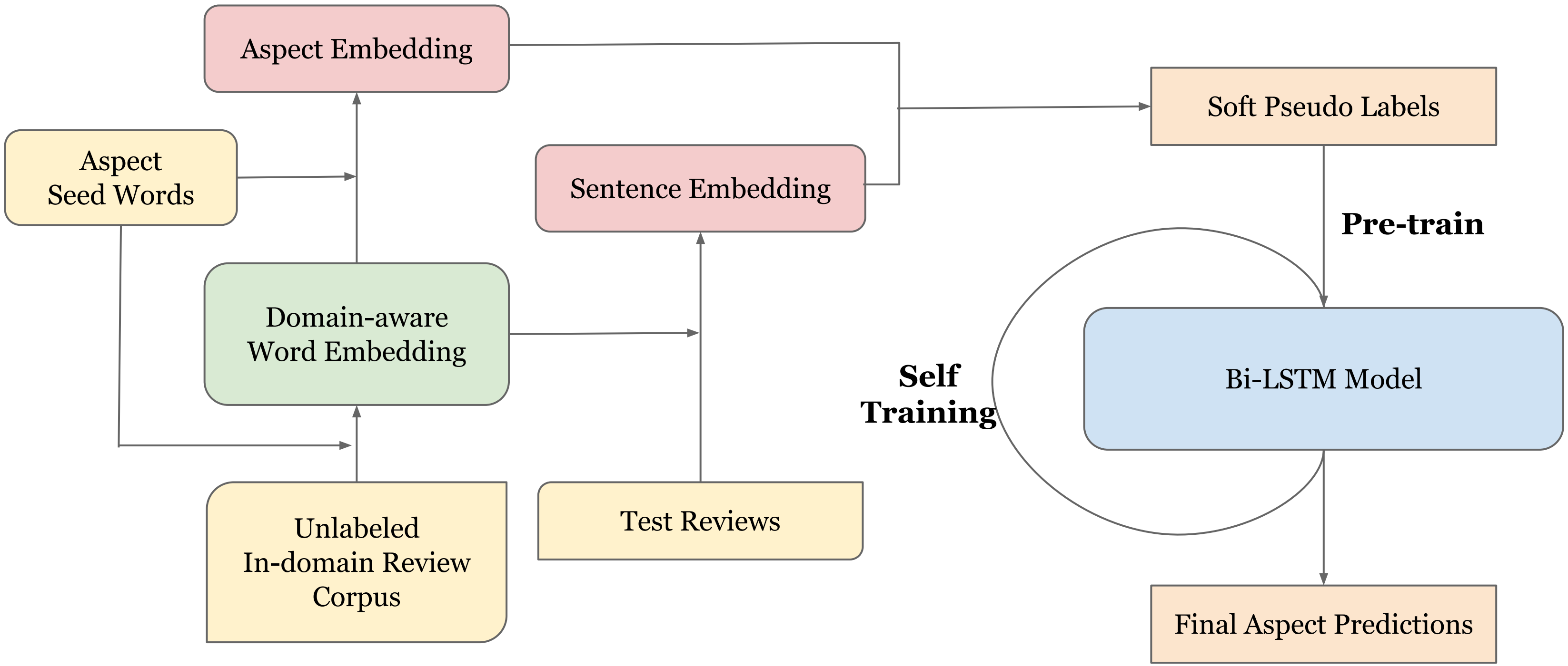
- Word embedding training corpus: 17,027 unlabeled reviews from Yelp Dataset Challenge.
- Evaluation dataset: the **restaurant** domain of SemEval-2016 and SemEval-2015.
  - 643 sentences after cleaning.
  - 5 Aspect: Location / Drinks / Food / Ambience / Service

Ex: (Training)
Burgerfi is shit. five_guys does it right. come eat a real burger here ! ! ! ! ! ! ! ! even a little burger and small_fry would fill up a hungry man ! ! bravo

Ex: (Evaluation)
We were then **charged** for their most **expensive sake** ($ 20+ per serving) when we in fact drank a sake of less than half that price . (Service)

Service not the **friendliest** to our `` large party '' ! (Service)

Framework



Aspect Embedding
Aspect
Seed Words
Sentence Embedding
Soft Pseudo Labels
Pre-train
Domain-aware
Word Embedding
Self
Training
Bi-LSTM Model
Unlabeled
In-domain Review
Corpus
Test Reviews
Final Aspect Predictions

Aspect Seed Words

Aspect	Seed Words
Location	street block river avenue
Drinks	beverage wines cocktail sake
Food	spicy sushi pizza tasty
Ambience	romantic room seating small
Service	tips manager waitress servers

Table 1. A typical seed word list for restaurant review corpus

CosSim(word_emb[‘spicy’], word_emb[‘pizza’]) = 0.1556
CosSim(word_emb[‘tasty’], word_emb[‘pizza’]) = 0.2489
CosSim(word_emb[‘spicy’], word_emb[‘tasty’]) = 0.4013

Location_n	street ...	Location_adj	convenient ...
Drinks_n	wines ...	Drinks_adj	iced ...
Food_n	pizza ...	Food_adj	tasty ...
Ambience_n	room ...	Ambience_adj	small ...
Service_n	tips ...	Service_adj	friendly ...

Aspect Seed Words

$$S_w = \sum_{a=1}^{k} \sum_{v \in S(a)} (x_v - \mu_a)(x_v - \mu_a)^T$$

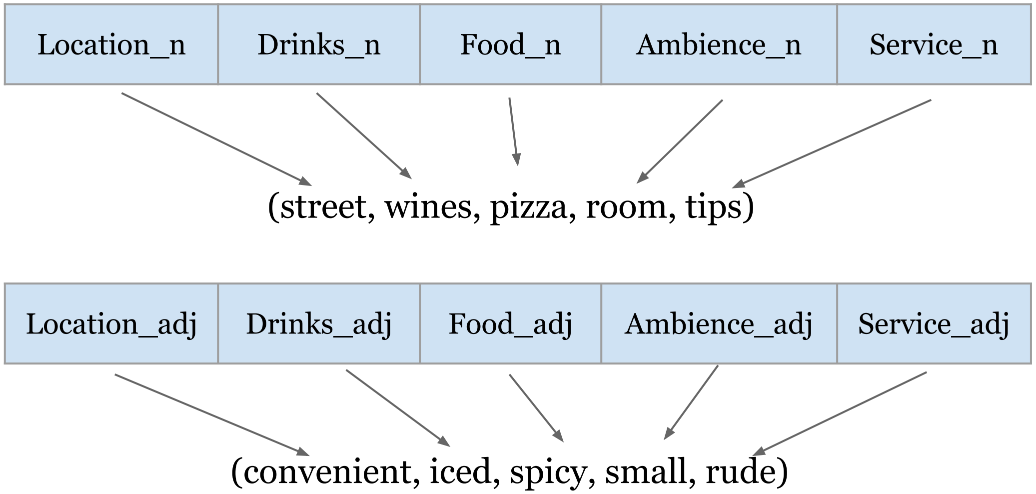
$$S_b = \sum_{a=1}^{k} |S(a)|(\mu_a - \mu)(\mu_a - \mu)^T$$

$$J = \frac{S_b}{S_w}$$

S_w: within aspect variance
S_b: between aspect variance
J: the larger the better

	S_w	J
5 Aspects	11.6723	0.2554
5 Aspects w/ n. and adj. seperated	9.4199	0.4931

Aspect-Discriminative Word Embedding



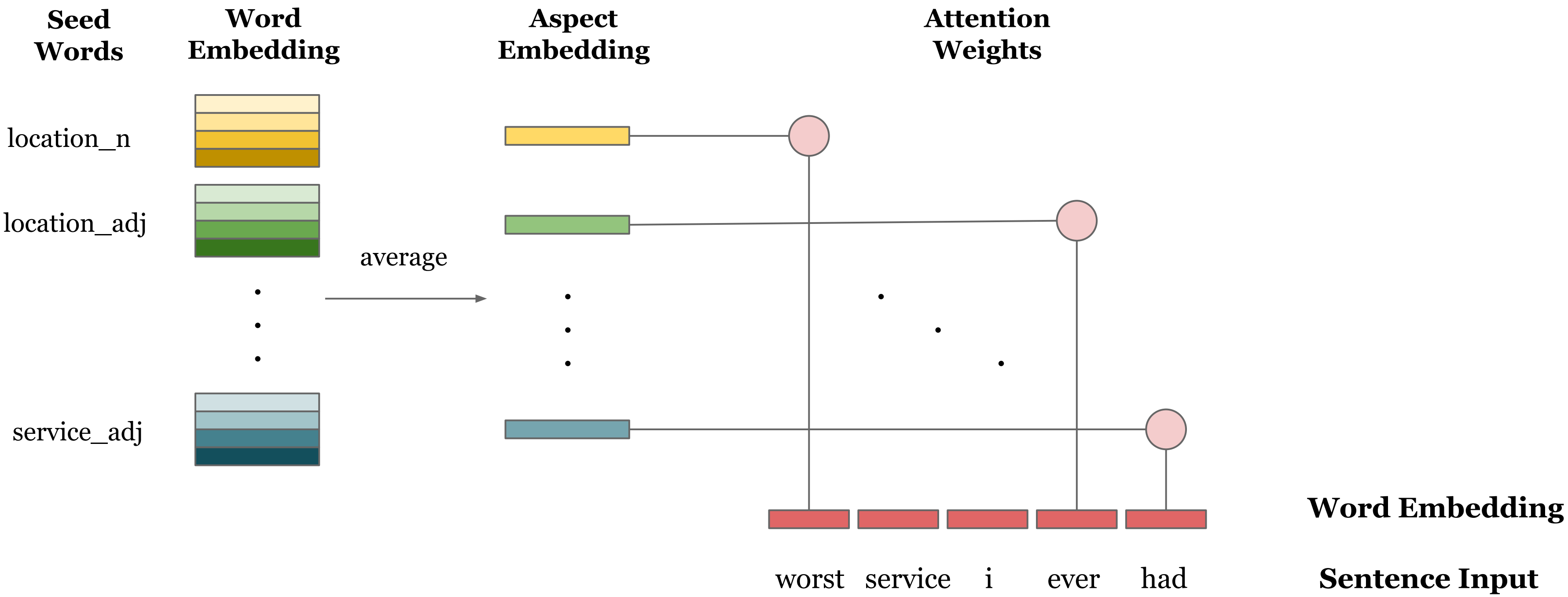
$$L_d = \sum_w \sum_v CosSim(x_w, x_v)$$

$$L = min(L_{contect} + \alpha L_d)$$

	S_b	$J=S_b/S_w$
Original Word2Vec	5.0907	0.4931
Constrained Word2Vec	5.1048	0.4982

Liu, Qian, et al. "Semantic structure-based word embedding by incorporating concept convergence and word divergence." *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*.

Pseudo Label Generation



Li, P., Guo, F., & Shang, J. (2020). User-Guided Aspect Classification for Domain-Specific Texts. arXiv preprint arXiv:2004.14555.

Pseudo Label Generation

$$Attention(word_i) = \max_{a \in Aspects} (CosSim(x_i, X_a))$$

$$X_{sentence} = \sum_{i \in sentence} w_i x_i$$

$$w_i = \frac{e^{Attention(i)}}{\sum_{i \in sentence} e^{Attention(i)}}$$

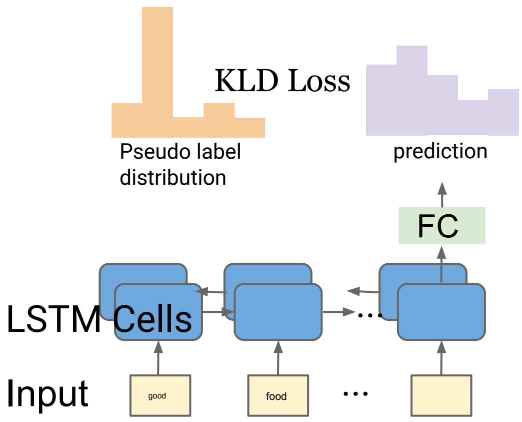
$$Pseudo\,Label(sentence) = argmax_{a \in Aspects}(CosSim(X_{sentenct}, X_a))$$

Refinement

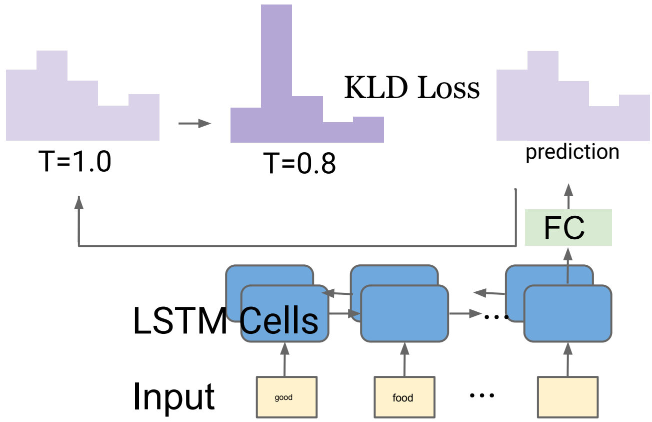
Purpose: Word level semantics->Sentence level/Sequential information
Model: Bi-directional LSTM/Pre-trained word embedding

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

1. Pre-train
Confidence score given by pseudo label as Soft-label.



2. Self-train
Refine model by recent high-confidence prediction.



Experiments

Model	Accuracy	Precision	Recall	Macro-F1
CosSim	61.43	50.12	50.26	42.31
ABAE	67.34	46.63	50.79	45.31
ARYA	81.34	62.05	58.38	56.24
JASen	83.83	64.73	72.95	66.28
Ours	**87.09**	**77.15**	**76.52**	**76.67**

CosSim: Aspect mean is the seed word average. Prediction is based on the CosSim between sentence and aspect mean.

ABAE: Autoencoder-based model

ARYA: User-Guided Aspect Classification for Domain-Specific Texts

JASen: Joint Aspect-Sentiment Topic Embedding

Ablation Study

- 5-Aspect: Without specification of adj/n
- w/ set expansion: Expand seed word by selecting Nearest Neighbor while calculating aspect embedding
- w/o Aspect-Discriminative Word Embedding: word embedding is original word2vec
- w/o self-training: An ablation model with only pre-training on LSTM

Model	Accuracy	Precision	Recall	Macro-F1
5-Aspect	86.00	76.53	64.63	67.65
w/ set expansion	85.23	75.22	64.74	67.58
w/o Aspect-Discri Word Embedding	**87.09**	75.45	67.36	69.97
w/o Self-training	86.31	74.74	66.78	69.19
Ours	**87.09**	**77.15**	**76.52**	**76.67**

Case Study 1

Input	Green	tea	creme	brulee	Is	a	must
CosSim	0.34	**0.66**	0.23	0.26	0.51	-	0.26
Nearest Aspect	Food_adj	**Drinks_adj**	Food_n	Food_n	Ambience_n	-	Food_adj

Ground Truth	Food
Pseudo Label	Drinks
Final Prediction	Food

Case Study 2

Input	Do	n't	leave	the	restaurant	without	it
CosSim	**0.34**	-	**0.29**	0.48	0.49	**0.26**	0.48
Nearest Aspect	**Service_n**	-	**Service_n**	Food_n	Ambience_n	**Service_n**	Food_n

Ground Truth	Food
Pseudo Label	Service
Final Prediction	Food

Failure attempts

- Training Word2vec with the constraint that the seed word embeddings from the same aspect should be similar
- Seed word expansion
- De-sentiment on adjective seed words