CYUT at IJCNLP-2017 Task 3: System Report for Review Opinion Diversification

Shih-Hung Wu [1], Su-Yu Chang [2], and Liang-Pu Chen[3]
[1, 2] Dept. of CSIE, Chaoyang University of Technology, Taichung, Taiwan (R.O.C)
[3]DSI, Institute for Information Industry, Taipei, Taiwan (R.O.C)

Abstract

Review Opinion Diversification (RevOpiD) 2017 is a shared task which is held in International Joint Conference on Natural Language Processing (IJCNLP). The shared task aims at selecting top-k reviews, as a summary, from a set of re-views. There are three subtasks in RevOpiD: helpfulness ranking, representativeness ranking, and exhaustive coverage ranking. This year, our team submitted runs by three models. We focus on ranking reviews based on the helpfulness of the reviews. In the first two models, we use linear regression with two different loss functions. First one is least squares, and second one is cross entropy. The third run is a random baseline. For both k=5 and k=10, our second model gets the best scores in the official evaluation metrics.

1 Introduction

This paper reports how our team participated the Review Opinion Diversification (RevOpiD) 2017[1] shared task held in International Joint Conference on Natural Language Processing (IJCNLP). The shared task aims at selecting top-k reviews from a set of Amazon online product reviews on three different aspects, which are corresponding to three subtasks in RevOpiD: helpfulness ranking, representativeness ranking, and exhaustive coverage ranking (Singh et al., 2017).

This year, for k=5 and k=10, our team submitted three runs each by three models. We focus on ranking reviews based on the helpfulness of the reviews. In the first two models, we use linear regression with two different loss functions. The third run is a random baseline.

The paper is organized as follows: Section 2 gives the basic thought of how we construct our system. Section 3 shows our system architecture. The result is discussed in section 4. The conclusion and future works is in section 5.

2 Methodology

Our system follows the general machine learning approach. 1. Prepare the training data, 2. Find the proper features, 3. Train a model, and 4. Evaluate the result.

After observing the training data a little bit, we found that there are many reviews with zero vote (e.g. helpful[0/0]), which means there is no one voting this review at all. We cannot tell whether the review is not helpful, so that nobody voted, or the review is too new so people had no opportunity to vote it before the data was gathered. Therefore, we decide to filter out all the reviews with zero vote in the training set. The data preprocessing helps to get a better training result on training set. The zero vote data cause a lot of training error, since the zero vote data will make a regression system to give very low weights on all features.

Our system used two kinds of features: the length of a review, and the numbers of words with certain part-of-speech (POS) in the review; based on our experience on Chinese online review helpfulness prediction. In our previous works, we found that the distribution of certain part-of-speech (POS) will affect the ranking of opinion (Hsieh et al., 2014). Traditionally speaking, verbs, nouns, and adjectives are grouped as content words. The more content words are involved, the more informative, so the more helpful, the review is.

We chose the linear regression model this year. Many previous works have shown that linear regression model can be used to predict the helpfulness (Wu et al., 2017).

Our optimization goal is to rank the helpfulness according to the helpful votes. The problem has been studied by several previous works and shows promising result that text analysis results can help

[1] https://sites.google.com/itbhu.ac.in/revopid-2017

Proceedings of the 8th International Joint Conference on Natural Language Processing, Shared Tasks, pages 134–137,
Taipei, Taiwan, November 27 – December 1, 2017.

helpfulness prediction (Zeng and Wu. 2013)(Zeng et al., 2014).

3 System Architecture

3.1 Data Preprocessing

During the pre-processing, our system filtered out the reviews with zero vote. There are 3,619,981 reviews in the Training data. After filtering out the zero vote reviews, there are only 1,215,671 reviews remaining in our training set. There are 2,404,310 zero-vote-reviews, which occupies about 66% of the original training set.

3.2 Features

Our system used four features: the first one is the length of a review. The second to fourth ones are the numbers of verbs (VB), nouns (NN), and adjectives (JJ) in the review. The POS of words in review is tagged by the tagging function of a python toolkit NLTK (Loper and Bird, 2002). The tag set is defined as Penn treebank (Santorini, 1990), shown in Table 1. Actually there are other tags that also verbs (VBD, VBG, VBN, VBP, VBZ), nouns (NNS, NNP, NNPS), and adjectives (JJR, JJS). Due to the time limitation, we do not count them in in our system. We believed that the proportion of each POS tag in the reviews should be similar.

3.3 The Linear Regression Model A

To implement the linear regression in model A, we use the Python Scikit-learn (Pedregosa et al., 2011) In this linear regression module, the training data is standardized by the fit_transform() function, and the loss function is Least squares. The test data is then ranked according to the helpfulness prediction of the regression model.

3.4 The Linear Regression Model B

The second model is implemented with the Google TensorFlow toolkit (Allaire et al., 2016). The training data is not standardized. The linear regression formula is as follows:

$$\text{Hypothesis} = (W*X) + b \qquad (1)$$

where X is the input data matrix. The weights W and the bias b are randomly initialized. The learning rate is 0.01. The optimizer is GradientDescentOptimizer. The training_epochs is 10,000. The loss function is the reduce_mean function, which is the average cross entropy of each training batch. The model is then used as our second model. The test data is then ranked according to the helpfulness prediction of the regression model.

	Tag	Description
1	CC	Coordinating conjunction
2	CD	Cardinal number
3	DT	Determiner
4	EX	Existential *there*
5	FW	Foreign word
6	IN	Preposition or subordinating conjunction
7	JJ	Adjective
8	JJR	Adjective, comparative
9	JJS	Adjective, superlative
10	LS	List item marker
11	MD	Modal
12	NN	Noun, singular or mass
13	NNS	Noun, plural
14	NNP	Proper noun, singular
15	NNPS	Proper noun, plural
16	PDT	Predeterminer
17	POS	Possessive ending
18	PRP	Personal pronoun
19	PRP$	Possessive pronoun
20	RB	Adverb
21	RBR	Adverb, comparative
22	RBS	Adverb, superlative
23	RP	Particle
24	SYM	Symbol
25	TO	*to*
26	UH	Interjection
27	VB	Verb, base form
28	VBD	Verb, past tense
29	VBG	Verb, gerund or present participle
30	VBN	Verb, past participle
31	VBP	Verb, non-3rd person singular present
32	VBZ	Verb, 3rd person singular present
33	WDT	Wh-determiner
34	WP	Wh-pronoun
35	WP$	Possessive wh-pronoun
36	WRB	Wh-adverb

Table 1: part-of-speech tags used in the Penn Treebank

4 Experiments

4.1 The Data Set

Data set is provide by the task organizer. The training, development and test data have been extracted and annotated from Amazon SNAP Review Dataset. (He and McAuley, 2016)

	For sub-task A	For subtask B					For subtask C	
METRICS LIST:	mth	cos_d	cos	cpr	a-dcg	wt	unwt	recall
CYUT1_A_5	0.71	0.83	0.84	0.7	4.28	504.18	14.31	0.71
CYUT2_A_5	**0.84**	**0.87**	**0.88**	0.7	**5.22**	**575.58**	**17.67**	**0.83**
CYUT3_A_5 (random baseline)	0.7	0.79	0.81	0.07	3.53	408.58	11.04	0.66
FAAD1_A_5	0.78	0.86	0.87	0.49	4.27	494.03	14.04	0.76
FAAD2_A_5	0.78	0.85	0.86	0.52	4.34	495.35	14.34	0.75
FAAD3_A_5	0.78	0.84	0.85	0.51	4.11	486.51	13.35	0.72
JUNLP_A_5	0.8	0.83	0.85	0.46	4.05	475.54	13.12	0.74
JUNLP_B_5	0.7	0.86	0.87	0.71	4.98	556.94	16.9	0.81
BASE_R_B_5	0.64	0.84	0.84	0.74	4.53	533.41	15.33	0.73
JUNLP_C_5	0.53	0.8	0.81	0.3	3.58	390.44	10.94	0.67

Table 2: Official Run Results of RevOpiD 2017 for k=5

	For sub-task A	For subtask B					For subtask C	
METRICS LIST:	mth	cos_d	cos	cpr	a-dcg	wt	unwt	recall
CYUT1_A_10	0.76	0.9	0.92	0.7	5.22	1280.6	36.53	0.89
CYUT2_A_10	**0.86**	0.91	0.92	**0.76**	**6.06**	**1517.1**	**45.79**	**0.95**
CYUT3_A_10 (random baseline)	0.75	0.89	0.9	0.14	4.48	1135.9	30.29	0.88
FAAD1_A_10	0.81	0.92	0.93	0.61	5.18	1325.2	37.54	0.89
FAAD2_A_10	0.84	0.91	0.92	0.65	5.2	1318.8	37.8	0.9
FAAD3_A_10	0.83	0.92	0.94	0.65	5.16	1317.5	36.8	0.92
JUNLP_A_10	0.84	0.91	0.92	0.59	5.04	1301.4	36.21	0.92
JUNLP_B_10	0.75	0.9	0.91	0.68	5.71	1384.6	41.03	0.91
JUNLP_C_10	0.73	0.88	0.9	0.39	4.46	1045.6	28.93	0.85

Table 3: Official Run Results of RevOpiD 2017 for k=10

4.2 The Official Evaluation Results

The official evaluation results is shown in Table 2 and 3. Our three runs are denoted as CYUT#_A_k for k=5 and k=10, # for 1, 2, and 3. The metric abbreviations are as follows (Singh et al., 2017):

For subtask A:

mth: The fraction of reviews included with more than half votes in favor.

For subtask B:

cos_d: discounted cosine similarity

cos: Cosine Similarity

cpr: cumulative proportionality (Dang and Croft, 2012)

a-dcg: Alpha-DCG (Clarke et al., 2008)

wt: weighted relevance

For subtask C:

unwt: unweighted relevance

recall: The fraction of opinions/columns covered by the top k ranked list.

4.3 Discussion

For k=5, the second run (**CYUT2_A_5**) gets the highest scores in seven of the eight official evaluation metrics. For k=10, the second run (**CYUT2_A_10**) gets the highest scores in six of the eight official evaluation metrics. This study shows that optimization the helpfulness (Subtask A) with cross entropy can also help exhaustive coverage (Subtask C), and help representativeness (Subtask B).

5 Conclusions and Future Works

Our team participated the RevOpiD 2017, focused on ranking reviews based on the helpfulness of the reviews. However, the result shows that it can also help on the exhaustive coverage, and representativeness. Our second linear regression model gets the highest scores in the official evaluation metrics for both k=5 and k=10.

We chose the linear regression model this year. There are still other machine learning models could be used in the future, such as deep neural networks. In deep learning paradigm, it is possible to bypass the feature engineering efforts. That is, we do not need to worry about which features are more useful.

Acknowledgments

This study is conducted under the "Online and Offline integrated Smart Commerce Platform (4/4)" of the Institute for Information Industry which is subsidized by the Ministry of Economic Affairs of the Republic of China. This study is also supported by the Ministry of Science under the grant numbers MOST106-2221-E-324-021-MY2.

References

JJ Allaire, Dirk Eddelbuettel, Nick Golding, and Yuan Tang. 2016. tensorflow: R Interface to TensorFlow. https://github.com/rstudio/tensorflow

Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '08). ACM, New York, NY, USA, 659-666. DOI: http://dx.doi.org/10.1145/1390334.1390446

Van Dang and W. Bruce Croft. 2012. Diversity by proportionality: an election-based approach to search result diversification. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '12). ACM, New York, NY, USA, 65-74. DOI: https://doi.org/10.1145/2348283.2348296

Ruining He and Julian McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In Proceedings of the 25th International Conference on World Wide Web (WWW '16). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 507-517. DOI: https://doi.org/10.1145/2872427.2883037

Hsien-You Hsieh, Vitaly Klyuev, Qiangfu Zhao, Shih-Hung Wu. 2014. *SVR-based outlier detection and its application to hotel ranking*. In proceedings of IEEE 6th International Conference on Awareness Science and Technology (iCAST-2014), Paris, France.

Edward Loper and Steven Bird. 2002. NLTK: the Natural Language Toolkit. In Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1 (ETMTNLP '02), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 63-70. DOI: https://doi.org/10.3115/1118108.1118117

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12(Oct): 2825−2830.

Beatrice Santorini. 1990. "Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision)", July.

Anil Kumar Singh, Avijit Thawani, Anubhav Gupta and Rajesh Kumar Mundotiya. Evaluating Opinion Summarization in Ranking. Proceedings of the 13th Asia Information Retrieval Societies Conference (AIRS 2017). Jeju island, Korea. November, 2017.

Anil Kumar Singh, Avijit Thawani, Mayank Panchal, Anubhav Gupta and Julian McAuley. Overview of the IJCNLP-2017 Shared Task on Review Opinion Diversification (RevOpiD-2017). In Proceedings of the IJCNLP-2017 Shared Tasks. Taipei, Taiwan. December, 2017.

Shih-Hung Wu, Yi-Hsiang Hsieh, Liang-Pu Chen, Ping-Che Yang, Liu Fanghuizhu. 2017. *Temporal Model of the Online Customer Review Helpfulness Prediction*. in Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Sydney, Australia, 31 July - 03 August.

Yi-Ching Zeng and Shih-Hung Wu. 2013. *Modeling the Helpful Opinion Mining of Online Consumer Reviews as a Three-class Classification Problem*. in Proceedings of the SocialNLP workshop, Nagoya, Japan, Oct. 14, 2013.

Yi-Ching Zeng, Tsun Ku, Shih-Hung Wu, Liang-Pu Chen, and Gwo-Dong Chen. 2014. *Modeling the Helpful Opinion Mining of Online Consumer Reviews as a Classification Problem*, International Journal of Computational Linguistics & Chinese Language Processing ,Vol. 19, No. 2, June.