Metamorphic Testing of Cross-Language Sentiment Analysis

Boyang

June 26, 2018

Contents

1 Test Data

Total have 46180 movies reviews.

Ranking	Number of Test Data	Percentage
Ranking 10	7353	15.92 %
Ramking 20	11209	24.27 %
Ranking 30	16223	35.13 %
Ranking 40	7663	16.59 %
Ranking 50	3732	8.08 %

2 Assessing Machine translation tool quality

2.1 Method

1. Compare correlation coefficient between Chinese sentiment analysis results and English sentiment analysis results by each

   (a) Using Google, Baidu, Yandex translation tools, translated original Chinese data to English data

   (b) Using same sentiment analysis tool analysis original chinese dataset and translated dataset

   (c) Calculate correlation coefficient between Chinese sentiment analysis results and English sentiment analysis results

(d) Compare correlation coefficient values. if value is bigger than others, we can say this translation tool, which use in original dataset to English dataset, can achieve better results than others.
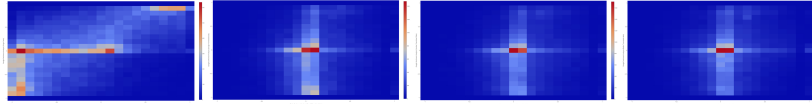
2.1.1 Result

- Base on Google sentiment analysis tool

	Google Score for Google translated data	Google Score for Yandex trans
Gooogle Score for origin data	0.512 (Pearson Correlations) p-value: 0.0	0.506 (Pearson Correlations) p
Google Score for origin data	0.381 (Kendall Correlations) p-value: 0.0	0.375 (Kendall Correlations) p
Google Score for origin data	0.504 (Spearman Correlations) p-value: 0.0	0.497 (Spearman Correlations)
Gooogle Score for origin data	0.512 (Point Biserial) p-value: 0.0	0.506 (Point Biserial) p-value:

- Google translation tool quality > Yandex translation tool quality > Baidu translation tool quality

- Base on Baidu sentiment analysis tool

	Baidu Positive Probability for Google translated data	Baidu P
Baidu Positive Probability for origin data	0.288 (Pearson Correlations) p-value: 0.0	0.280 (P
Baidu Positive Probability for origin data	0.188 (Kendall Correlations) p-value: 0.0	0.174 (P
Baidu Positive Probability for origin data	0.271 (Spearman Correlations) p-value: 0.0	0.249 (S
Baidu Positive Probability for origin data	0.288 (Point Biserial) p-value: 0.0	0.280 (P

- Google translation tool quality > Yandex translation tool quality > Baidu translation tool quality

1. Compare correlation coefficient between sentiment analysis results and user rating

   (a) Divide the sentiment analysis scores between [-1,1] into 5 regions, which are 10, 20, 30, 40, 50. Each interval is 0.4.

regions	scope
10	[-1, -0.6]
20	(-0.6, -0.2]
30	(-0.2, 0.2)
40	[0.2, 0.6)
50	[0.6, 1]

1. calculate correlation coefficient between regions and user rating

- Base on Google sentiment analysis tool

	Google score for Google Translated data	Google score for Yandex Translated data	
ranking	0.3639 (pearson Correlation) p-value: 0.0	0.3621 (pearson correlation) p-value: 0.0	
ranking	0.3645 (spearman correlation) p-value: 0.0	0.3623 (spearman correlation) p-value: 0.0	
ranking	0.3689 (point biserial correlation) p-value: 0.0	0.3621 (point biserial correlation) p-value: 0.0	
ranking	0.3009 (kendall correlation) p-value: 0.0	0.2999 (kendall correlation) p-value: 0.0	

- Base on Baidu sentiment analysis tool

	Baidu sentiment score (Google standard) for Google Translated data	Baidu sentiment score (Google standard) for Yandex Translated data	Baidu sentiment score (Google standard) for Baidu Translated data
ranking	0.1638 (pearson correlation) p-value: 0.0	0.1645 (pearson correlation) p-value: 0.0	0.1611 (pearson correlation) p-value: 0.0
ranking	0.1687 (spearman correlation) p-value: 0.0	0.1682 (spearman correlation) p-value: 0.0	0.1624 (spearman correlation) p-value: 0.0
ranking	0.1638 (point biserial correlation) p-value: 0.0	0.1645 (point biserial correlation) p-value: 0.0	0.1611 (point biserial correlation) p-value: 0.0
ranking	0.1455 (kendall correlation) p-value: 0.0	0.1447 (kendall correlation) p-value: 0.0	0.1394 (kendall correlation) p-value: 0.0

1. draw heatmap

   (a) Divide the sentiment analysis scores between [-1, 1] into 20 regions, 0, 1, 2, 3 . . . 18, 19. Each interval is 0.1

   (b) Draw the heatmaps between the user rating (i.e., 10, 20, , 50) and sentiment analysis scores (20 subregions).