Results

August 14, 2018

1 Tables of Friedman, Bonferroni-Dunn, Holm, Hochberg and Hommel Tests

Table 1: Average Rankings of the algorithms

Algorithm	Ranking
Word2vec	4.035714285714286
BOW-TFIDF	3.642857142857143
Simple encoder-decoder	4.0
Google universal encoder	3.285714285714285
Skip-Thoughts	1.857142857142857
InferSent	4.178571428571428

Friedman statistic considering reduction performance (distributed according to chi-square with 5 degrees of freedom: 15.051020408163254. P-value computed by Friedman Test: 0.01014654573783258.

Iman and Davenport statistic considering reduction performance (distributed according to F-distribution with 5 and 65 degrees of freedom: 3.5608170844939613.

P-value computed by Iman and Daveport Test: 0.006595925370070002.

Table 2: Holm / Hochberg Table for $\alpha = 0.05$

i	algorithm	$z = (R_0 - R_i)/SE$	p	Holm/Hochberg/Hommel
5	InferSent	3.2829957697946845	0.00102710206925608	0.01
4	Word2vec	3.080965260884243	0.002063072007576214	0.0125
3	Simple encoder-decoder	3.0304576336566327	0.0024418344202768298	0.016666666666666666
2	BOW-TFIDF	2.5253813613805267	0.011557279664325225	0.025
1	Google universal encoder	2.0203050891044207	0.043351751260862804	0.05

Bonferroni-Dunn's procedure rejects those hypotheses that have a p-value ≤ 0.01.
Hochberg's procedure rejects those hypotheses that have a p-value ≤ 0.05.
Hommel's procedure rejects all hypotheses.

Table 3: Holm / Hochberg Table for $\alpha = 0.10$

i	algorithm	$z = (R_0 - R_i)/SE$	p	Holm/Hochberg/Hommel
5	InferSent	3.2829957697946845	0.00102710206925608	0.02
4	Word2vec	3.080965260884243	0.002063072007576214	0.025
3	Simple encoder-decoder	3.0304576336566327	0.0024418344202768298	0.03333333333333333
2	BOW-TFIDF	2.5253813613805267	0.011557279664325225	0.05
1	Google universal encoder	2.0203050891044207	0.043351751260862804	0.1

Bonferroni-Dunn's procedure rejects those hypotheses that have a p-value ≤ 0.02.
Hochberg's procedure rejects those hypotheses that have a p-value ≤ 0.1.
Hommel's procedure rejects all hypotheses.

Table 4: Adjusted p-values

i	algorithm	unadjusted p	p_{Bonf}	p_{Holm}	p_{Hoch}	p_{Homm}
1	InferSent	0.00102710206925608	0.0051355103462804004	0.0051355103462804004	0.0051355103462804004	0.00410840827702432
2	Word2vec	0.002063072007576214	0.010316536003788106	0.008253228803030485	0.00732550326083049	0.006189921602272864 5
3	Simple encoder-decoder	0.0024418344202768298	0.012209172101384148	0.008253228803030485	0.00732550326083049	0.00732550326083049
4	BOW-TFIDF	0.011557279664325225	0.057786398321626124	0.02311455932865045	0.02311455932865045	0.02311455932865045
5	Google universal encoder	0.043351751260862804	0.21675875630431402	0.043351751260862804	0.043351751260862804	0.043351751260862804

Table 5: Holm / Shaffer Table for $\alpha = 0.05$

i	algorithms	$z = (R_0 - R_i)/SE$	p	Holm	Shaffer
15	Skip-Thoughts vs. InferSent	3.2829957697946845	0.00102710206925608	0.0033333333333333335	0.0033333333333333335
14	Word2vec vs. Skip-Thoughts	3.080965260884243	0.0020633072007576214	0.0035714285714285718	0.005
13	Simple encoder-decoder vs. Skip-Thoughts	3.0304576336566327	0.0024418344202768298	0.0038461538461538464	0.005
12	BOW-TFIDF vs. Skip-Thoughts	2.5253813613805267	0.011557279664325225	0.004166666666666667	0.005
11	Google universal encoder vs. Skip-Thoughts	2.0203050891044207	0.043351751260862804	0.004545454545454546	0.005
10	Google universal encoder vs. InferSent	1.2626906806902634	0.20670036450499135	0.005	0.005
9	Word2vec vs. Google universal encoder	1.0606601717798219	0.28884436634648464	0.005555555555555556	0.005555555555555556
8	Simple encoder-decoder vs. Google universal encoder	1.0101525445522115	0.312422211242693	0.00625	0.00625
7	BOW-TFIDF vs. InferSent	0.7576144084141572	0.44868185197701566	0.0071428571428571435	0.0071428571428571435
6	Word2vec vs. BOW-TFIDF	0.5555838995037158	0.5784953405056155	0.008333333333333333	0.008333333333333333
5	BOW-TFIDF vs. Google universal encoder	0.5050762722761061	0.6135052661279103	0.01	0.01
4	BOW-TFIDF vs. Simple encoder-decoder	0.505076272761054	0.6135052661279108	0.0125	0.0125
3	Simple encoder-decoder vs. InferSent	0.25253813613805176	0.8006251407096018	0.016666666666666666	0.016666666666666666
2	Word2vec vs. InferSent	0.20203050891044141	0.8398928732712694	0.025	0.025
1	Word2vec vs. Simple encoder-decoder	0.050507627227610354	0.9597178715153033	0.05	0.05

Nemenyi's procedure rejects those hypotheses that have a p-value $\leq 0.0033333333333333335$.
Holm's procedure rejects those hypotheses that have a p-value $\leq 0.004166666666666667$.
Shaffer's procedure rejects those hypotheses that have a p-value $\leq 0.0033333333333333335$.
Bergmann's procedure rejects these hypotheses:

- Word2vec vs. Skip-Thoughts
- Simple encoder-decoder vs. Skip-Thoughts
- Skip-Thoughts vs. InferSent

Nemenyi's procedure rejects those hypotheses that have a p-value $\leq 0.006666666666666667$.
Holm's procedure rejects those hypotheses that have a p-value $\leq 0.008333333333333333$.
Shaffer's procedure rejects those hypotheses that have a p-value $\leq 0.006666666666666667$.
Bergmann's procedure rejects these hypotheses:

- Word2vec vs. Skip-Thoughts
- BOW-TFIDF vs. Skip-Thoughts

Table 6: Holm / Shaffer Table for $\alpha = 0.10$

i	algorithms	$z = (R_0 - R_i)/SE$	p	Holm	Shaffer
15	Skip-Thoughts vs. InferSent	3.2829957697946845	0.00102710206925608	0.006666666666666667	0.006666666666666667
14	Word2vec vs. Skip-Thoughts	3.080965260884243	0.0020633072007576214	0.007142857142857143 5	0.01
13	Simple encoder-decoder vs. Skip-Thoughts	3.0304576336566327	0.0024418344202768298	0.007692307692307693	0.01
12	BOW-TFIDF vs. Skip-Thoughts	2.5253813613805267	0.011557279664325225	0.008333333333333333	0.01
11	Google universal encoder vs. Skip-Thoughts	2.0203050891044207	0.043351751260862804	0.009090909090909092	0.01
10	Google universal encoder vs. InferSent	1.2626906806902634	0.20670036450499135	0.01	0.01
9	Word2vec vs. Google universal encoder	1.0606601717798219	0.28884436634648464	0.011111111111111112	0.011111111111111112
8	Simple encoder-decoder vs. Google universal encoder	1.0101525445522115	0.312422211242693	0.0125	0.0125
7	BOW-TFIDF vs. InferSent	0.7576144084141572	0.44868185197701566	0.014285714285714287	0.014285714285714287
6	Word2vec vs. BOW-TFIDF	0.5555838995037158	0.5784953405056155	0.016666666666666666	0.016666666666666666
5	BOW-TFIDF vs. Google universal encoder	0.5050762722761061	0.6135052661279103	0.02	0.02
4	BOW-TFIDF vs. Simple encoder-decoder	0.5050762722761054	0.6135052661279108	0.025	0.025
3	Simple encoder-decoder vs. InferSent	0.25253813613805176	0.8006251407096018	0.03333333333333333	0.03333333333333333
2	Word2vec vs. InferSent	0.20203050891044141	0.8398928732712694	0.05	0.05
1	Word2vec vs. Simple encoder-decoder	0.050507627227610354	0.9597178715153033	0.1	0.1

- Simple encoder-decoder vs. Skip-Thoughts
- Skip-Thoughts vs. InferSent

Table 7: Adjusted p-values

i	hypothesis	unadjusted p	p_{Neme}	p_{Holm}	p_{Shaf}	p_{Berg}
1	Skip-Thoughts vs .InferSent	0.00102710206925608	0.0154065310388412	0.0154065310388412	0.0154065310388412	0.0154065310388412
2	Word2vec vs .Skip-Thoughts	0.0020633072007576214	0.03094960801136432	0.0288863008106067	0.020633072007576213	0.020633072007576213
3	Simple encoder-decoder vs .Skip-Thoughts	0.0024418344202768298	0.03662751630415244 5	0.03174384746359879	0.024418344202768297	0.020633072007576213
4	BOW-TFIDF vs .Skip-Thoughts	0.011557279664325225	0.17335919496487837	0.1386873559719027	0.11557279664325225	0.08090095765027658
5	Google universal encoder vs .Skip-Thoughts	0.043351751260862804	0.6502762689129421	0.47686926386949086	0.43351751260862803	0.3034622588260396
6	Google universal encoder vs .InferSent	0.20670036450499135	3.1005054675748704	2.0670036450499136	2.0670036450499136	2.0670036450499136
7	Word2vec vs .Google universal encoder	0.28884436634648464	4.332665495197269	2.599599297118362	2.0670036450499136	2.0670036450499136
8	Simple encoder-decoder vs .Google universal encoder	0.312422211242693	4.686333168640394	2.599599297118362	2.186955478698851	2.0670036450499136
9	BOW-TFIDF vs .InferSent	0.44868185197701566	6.730227779655235	3.1407729638391095	3.1407729638391095	2.692091111862094
10	Word2vec vs .BOW-TFIDF	0.5784953405056155	8.677430107584232	3.4709720430336928	3.4709720430336928	2.692091111862094
11	BOW-TFIDF vs .Google universal encoder	0.6135052661279103	9.202578991918655	3.4709720430336928	3.4709720430336928	2.692091111862094
12	BOW-TFIDF vs .Simple encoder-decoder	0.6135052661279108	9.202578991918662	3.4709720430336928	3.4709720430336928	2.692091111862094
13	Simple encoder-decoder vs .InferSent	0.8006251407096018	12.009377110644026	3.4709720430336928	3.4709720430336928	2.692091111862094
14	Word2vec vs .InferSent	0.8398928732712694	12.59839309906904	3.4709720430336928	3.4709720430336928	2.692091111862094
15	Word2vec vs .Simple encoder-decoder	0.9597178715153033	14.395768072729549	3.4709720430336928	3.4709720430336928	2.692091111862094