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	4.178571428571428
BOW-TFIDF	3.178571428571429
Simple encoder-decoder	3.392857142857143
Google universal encoder	4.07142857142857
Skip-Thoughts	2.0357142857142856
InferSent	4.142857142857142

Friedman statistic considering reduction performance (distributed according to chi-square with 5 degrees of freedom: 13.836734693877474. P-value computed by Friedman Test: 0.016680443454981875.

Iman and Davenport statistic considering reduction performance (distributed according to F-distribution with 5 and 65 degrees of freedom: 3.2027616279069546.

P-value computed by Iman and Daveport Test: 0.012033882162591153.

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

i	algorithm	$z = (R_0 - R_i)/SE$	p	Holm/Hochberg/Hommel
5	Word2vec	3.0304576336566313	0.002441834420276842	0.01
4	InferSent	2.979950006429021	0.0028829543049806883	0.0125
3	Google universal encoder	2.878934751973799	0.003990208794973573	0.016666666666666666
2	Simple encoder-decoder	1.9192898346492004	0.05494766378708745	0.025
1	BOW-TFIDF	1.6162440712835375	0.10604155370017063	0.05

Bonferroni-Dunn's procedure rejects those hypotheses that have a p-value ≤ 0.01.
Holm's procedure rejects those hypotheses that have a p-value ≤ 0.025.
Hochberg's procedure rejects those hypotheses that have a p-value $\leq 0.016666666666666666$.
Hommel's procedure rejects those hypotheses that have a p-value ≤ 0.025.

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

i	algorithm	$z = (R_0 - R_i)/SE$	p	Holm/Hochberg/Hommel
5	Word2vec	3.0304576336566313	0.002441834420276842	0.02
4	InferSent	2.979950006429021	0.0028829543049806883	0.025
3	Google universal encoder	2.878934751973799	0.003990208794973573	0.03333333333333333
2	Simple encoder-decoder	1.9192898346492004	0.05494766378708745	0.05
1	BOW-TFIDF	1.6162440712835375	0.10604155370017063	0.1

Bonferroni-Dunn's procedure rejects those hypotheses that have a p-value ≤ 0.02.
Holm's procedure rejects those hypotheses that have a p-value ≤ 0.05.
Hochberg's procedure rejects those hypotheses that have a p-value ≤ 0.03333333333333333.
Hommel's procedure rejects those hypotheses that have a p-value ≤ 0.025.
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 $\leq 0.0038461538461538464$.
Shaffer's procedure rejects those hypotheses that have a p-value $\leq 0.0033333333333333335$.
Bergmann's procedure rejects these hypotheses:

- Word2vec vs. Skip-Thoughts

Table 4: Adjusted p-values

i	algorithm	unadjusted p	p_{Bonf}	p_{Holm}	p_{Hoch}	p_{Homm}
1	Word2vec	0.002441834420276842	0.012209172101384209	0.012209172101384209	0.011531817219922753	0.007980417589
2	InferSent	0.0028829543049806883	0.014414771524903442	0.012209172101384209	0.011531817219922753	0.008648862914
3	Google universal encoder	0.003990208794973573	0.01995104397486787	0.012209172101384209	0.01197062638492072	0.011970626384
4	Simple encoder-decoder	0.05494766378708745	0.27473831893543726	0.1098953275741749	0.10604155370017063	0.106041553700
5	BOW-TFIDF	0.10604155370017063	0.5302077685008532	0.1098953275741749	0.10604155370017063	0.106041553700

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

i	algorithms	$z = (R_0 - R_i)/SE$	p	Holm	Shaffer
15	Word2vec vs. Skip-Thoughts	3.0304576336566313	0.002441834420276842	0.0033333333333333335	0.00333333333
14	Skip-Thoughts vs. InferSent	2.979950006429021	0.0028829543049806883	0.003571428571428571 8	0.005
13	Google universal encoder vs. Skip-Thoughts	2.878934751973799	0.003990208794973573	0.0038461538461538464	0.005
12	Simple encoder-decoder vs. Skip-Thoughts	1.9192898346492004	0.05494766378708745	0.004166666666666667	0.005
11	BOW-TFIDF vs. Skip-Thoughts	1.6162440712835375	0.10604155370017063	0.004545454545454546	0.005
10	Word2vec vs. BOW-TFIDF	1.4142135623730936	0.1572992070502856	0.005	0.005
9	BOW-TFIDF vs. InferSent	1.3637059351454834	0.1726601353375736	0.005555555555555556	0.00555555555
8	BOW-TFIDF vs. Google universal encoder	1.2626906806902614	0.20670036450499205	0.00625	0.00625
7	Word2vec vs. Simple encoder-decoder	1.111167799007431	0.2664961289575324	0.007142857142857143 5	0.007142857142
6	Simple encoder-decoder vs. InferSent	1.0606601717798205	0.28884436634648514	0.008333333333333333	0.008333333333
5	Simple encoder-decoder vs. Google universal encoder	0.9596449173245987	0.3372339544842751	0.01	0.01
4	BOW-TFIDF vs. Simple encoder-decoder	0.30304576336566275	0.7618549862187104	0.0125	0.0125
3	Word2vec vs. Google universal encoder	0.15152288168283232	0.8795632622484182	0.016666666666666666	0.016666666666
2	Google universal encoder vs. InferSent	0.10101525445522197	0.9195383508138878	0.025	0.025
1	Word2vec vs. InferSent	0.050507627227610354	0.9597178715153033	0.05	0.05

- Google universal encoder vs. Skip-Thoughts
- Skip-Thoughts vs. InferSent

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

i	algorithms	$z = (R_0 - R_i)/SE$	p	Holm	Shaffer
15	Word2vec vs. Skip-Thoughts	3.0304576336566313	0.002441834420276842	0.006666666666666667	0.006666666666666667
14	Skip-Thoughts vs. InferSent	2.979950006429021	0.0028829543049806883	0.007142857142857143	0.01
13	Google universal encoder vs. Skip-Thoughts	2.878934751973799	0.003990208794973573	0.007692307692307693	0.01
12	Simple encoder-decoder vs. Skip-Thoughts	1.9192898346492004	0.05494766378708745	0.008333333333333333	0.01
11	BOW-TFIDF vs. Skip-Thoughts	1.6162440712835375	0.10604155370017063	0.009090909090909092	0.01
10	Word2vec vs. BOW-TFIDF	1.4142135623730936	0.1572992070502856	0.01	0.01
9	BOW-TFIDF vs. InferSent	1.3637059351454834	0.1726601353375736	0.011111111111111112	0.011111111111111112
8	BOW-TFIDF vs. Google universal encoder	1.2626906806902614	0.20670036450499205	0.0125	0.0125
7	Word2vec vs. Simple encoder-decoder	1.111167799007431	0.2664961289575324	0.014285714285714287	0.014285714285714287
6	Simple encoder-decoder vs. InferSent	1.0606601717798205	0.28884436634648514	0.016666666666666666	0.016666666666666666
5	Simple encoder-decoder vs. Google universal encoder	0.9596449173245987	0.3372339544842751	0.02	0.02
4	BOW-TFIDF vs. Simple encoder-decoder	0.30304576336566275	0.7618549862187104	0.025	0.025
3	Word2vec vs. Google universal encoder	0.15152288168283232	0.8795632622484182	0.03333333333333333	0.03333333333333333
2	Google universal encoder vs. InferSent	0.10101525445522197	0.9195383508138878	0.05	0.05
1	Word2vec vs. InferSent	0.050507627227610354	0.9597178715153033	0.1	0.1

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
- Google universal encoder 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	Word2vec vs .Skip-Thoughts	0.002441834420276842	0.036627516304152626	0.036627516304152626	0.036627516304152626	0.036627516304152626
2	Skip-Thoughts vs .InferSent	0.0028829543049806883	0.04324431457471033	0.04036136026972964	0.036627516304152626	0.036627516304152626
3	Google universal encoder vs .Skip-Thoughts	0.003990208794973573	0.0598531319246036	0.051872714334656456	0.03990208794973574	0.036627516304152626
4	Simple encoder-decoder vs .Skip-Thoughts	0.05494766378708745	0.8242149568063117	0.6593719654450494	0.5494766378708745	0.3846336465096122
5	BOW-TFIDF vs .Skip-Thoughts	0.10604155370017063	1.5906233055025594	1.166457090701877	1.0604155370017063	0.7422908759011944
6	Word2vec vs .BOW-TFIDF	0.1572992070502856	2.3594881057542842	1.5729920705028562	1.5729920705028562	1.5729920705028562
7	.BOW-TFIDF vs .InferSent	0.1726601353375736	2.589902030063604	1.5729920705028562	1.5729920705028562	1.5729920705028562
8	BOW-TFIDF vs .Google universal encoder	0.20670036450499205	3.1005054675748807	1.6536029160399364	1.5729920705028562	1.5729920705028562
9	Word2vec vs .Simple encoder-decoder	0.2664961289575324	3.997441934362986	1.8654729027027268	1.8654729027027268	1.5989767737451943
10	Simple encoder-decoder vs .InferSent	0.28884436634648514	4.332665495197277	1.8654729027027268	1.8654729027027268	1.5989767737451943
11	Simple encoder-decoder vs .Google universal encoder	0.3372339544842751	5.058509317264127	1.8654729027027268	1.8654729027027268	1.5989767737451943
12	BOW-TFIDF vs .Simple encoder-decoder	0.7618549862187104	11.427824793280656	3.0474199448748416	3.0474199448748416	3.0474199448748416
13	Word2vec vs .Google universal encoder	0.8795632622484182	13.193448933726273	3.0474199448748416	3.0474199448748416	3.0474199448748416
14	Google universal encoder vs .InferSent	0.9195383508138878	13.793075262208317	3.0474199448748416	3.0474199448748416	3.0474199448748416
15	Word2vec vs .InferSent	0.9597178715153033	14.395768072729549	3.0474199448748416	3.0474199448748416	3.0474199448748416