Validation set size= 7996 captions

4 reference captions per candidate

Vocabulary= 6839 words

| Individual Metric Correlation | Pearson | Spearman | Kendall |
|---|---|---|---|
| BLUE | 0.170 | 0.251 | 0.191 |
| ROGUE-L | 0.290 | 0.281 | 0.213 |
| METEOR | 0.411 | 0.447 | 0.341 |
| CIDER | 0.391 | 0.450 | 0.342 |
| SPICE | 0.440 | 0.453 | 0.349 |
| WMD | 0.405 | 0.431 | 0.328 |
| SPIDER-baseline | 0.415 | 0.468 | 0.357 |
| MCS-baseline | 0.423 | 0.471 | 0.360 |
| NNEval (old features) | | | |
| NNEval (new) | 0.536 | 0.535 | **0.413** |

Word Embeddings (we did not fine-tune the word embeddings)

| Name | Dimension | Corpus | Vocabulary |
|---|---|---|---|
| Glove_6B_50d | 50 | Wikipedia +Gigaword(6B) | 400K |
| Glove_42B_300d | 300 | Common Crawl (42B) | 1.9M |
| Glove_840B_300d | 300 | Common Crawl (840B) | 2.2M |
| Word2vec.Google_300d | 300 | Google News(100B) | 3M |
| Fast_text_300d | 300 | Wikipedia | 2.5M |

Semantic similarity using mean of words

*Results of **semantic sentence embedding**.*

| Name | Pearson | Spearman | Kendall |
|---|---|---|---|
| Glove_6B_50d | 0.349 | 0.361 | 0.273 |
| Glove_42B_300d | 0.433 | 0.437 | 0.332 |
| Glove_840B_300d | 0.448 | 0.442 | 0.337 |
| Word2vec.Google_300d | 0.474 | 0.458 | 0.349 |
| Fast_text_300d | 0.486 | 0.478 | **0.364** |

For syntactic features, we use the dependency tree parse that parses the sentence into a tree structure. Using that we evaluate the headword chains. These headword chains are then used to compute the similarity. We use POS chain lengths 2 to 4, whereas we also use the dependency and lemma variant HWCM chain length 2 because it gives a good correlation and perform better than BLEU and ROGUE-L

Results of Syntactic Features

|  | HWCM(POS) | HWCM(Lemma) | HWCM(Dep) |
|---|---|---|---|
| Max Length/Depth | Kendall Correlation |  |  |
| 1 | 0.083 | 0.089 | 0.292 |
| 2 | 0.075 | 0.040 | 0.297 |
| 3 | 0.078 | 0.026 | 0.293 |
| 4 | 0.085 | 0.019 | 0.291 |