Table 1: The nearest neighbours for the words "yale", "seahawks" and "eminem" based on the IN-IN and the IN-OUT vector cosine similarities. The IN-IN cosine similarities are high for words that are similar by function or type (*typical*), and the IN-OUT similarities are high between words that co-occur in the same query or document frequently (*topical*).

yale		seahawks		eminem	
IN-IN	IN-OUT	IN-IN	IN-OUT	IN-IN	IN-OUT
yale	yale	seahawks	seahawks	eminem	eminem
harvard	faculty	49ers	highlights	rihanna	rap
nyu	alumni	broncos	jerseys	ludacris	featuring
cornell	orientation	packers	tshirts	kanye	tracklist
tulane	haven	nfl	seattle	beyonce	diss
tufts	graduate	steelers	hats	2pac	performs

Also, the document embeddings can be pre-computed which is important for runtime efficiency. We only need to sum the score contributions across the query terms at the time of ranking.

As previously mentioned, the word2vec model contains two separate embedding spaces (IN and OUT) which gives us at least two variants of the DESM, corresponding to retrieval in the IN-OUT space or the IN-IN space[2].

$$DESM_{IN-OUT}(Q, D) = \frac{1}{|Q|} \sum_{q_i \in Q} \frac{q_{IN,i}^T \overline{D_{OUT}}}{\|q_{IN,i}\| \|\overline{D_{OUT}}\|} \quad (3)$$

$$DESM_{IN-IN}(Q, D) = \frac{1}{|Q|} \sum_{q_i \in Q} \frac{q_{IN,i}^T \overline{D_{IN}}}{\|q_{IN,i}\| \|\overline{D_{IN}}\|} \quad (4)$$

We expect the $DESM_{IN-OUT}$ to behave differently than the $DESM_{IN-IN}$ because of the difference in their notions of word relatedness as shown in Table 1.

One of the challenges of the embedding models is that they can only be applied to a fixed size vocabulary. We leave the exploration of possible strategies to deal with out-of-vocab (OOV) words for future investigation. In this paper, all the OOV words are ignored for computing the DESM score, but not for computing the TF-IDF feature, a potential advantage for the latter.

3. EXPERIMENTS

We train a *Continuous Bag-of-Words* (CBOW) model on a query corpus consisting of 618,644,170 queries and a vocabulary size of 2,748,230 words. The queries are sampled from Bing's large scale search logs from the period of August 19, 2014 to August 25, 2014. We repeat all our experiments using another CBOW model trained on a corpus of document body text with 341,787,174 distinct sentences sampled from the Bing search index and a corresponding vocabulary size of 5,108,278 words. Empirical results for both the models are presented in Table 2. Although our results are based exclusively on the CBOW model, the proposed methodology should be applicable to vectors produced by the *Skip-Gram* model, as both produce qualitatively and quantitatively similar embeddings.

We compare the retrieval performance of DESM against BM25 [6], a traditional count-based method, and Latent Semantic Analysis (LSA) [1], a traditional vector-based method. For the BM25 baseline we use the values of 1.7 for the k_1 and 0.95 for the b parameters based on a parameter sweep on a validation set. The LSA model is trained on the body text of 366,470 randomly sampled documents from Bing's index with a vocabulary size of 480,608 words. The evaluation set consists of 7,741 queries randomly sampled from

[2]It is also possible to define $DESM_{OUT-OUT}$ and $DESM_{OUT-IN}$, but we expect them to behave similar to $DESM_{IN-IN}$ and $DESM_{IN-OUT}$, respectively.

Table 2: The $DESM_{IN-OUT}$ performs significantly better than both the BM25 and the LSA baselines, as well as the $DESM_{IN-IN}$ on NDCG computed at positions three and ten. Also, the DESMs using embeddings trained on the query corpus performs better than if trained on document body text. The highest NDCG values for every column is highlighted in bold and all the statistically significant (p < 0.05) differences over the BM25 baseline are marked with the asterisk (*).

	NDCG@3	NDCG@10
BM25	29.14	44.77
LSA	28.25*	44.24*
DESM (IN-IN, trained on body text)	29.59	45.51*
DESM (IN-IN, trained on queries)	29.72	46.36*
DESM (IN-OUT, trained on body text)	30.32*	46.57*
DESM (IN-OUT, trained on queries)	**31.14***	**47.89***

Bing's query logs from the period of October, 2014 to December, 2014. For each sampled query, a set of candidate documents is constructed by retrieving the top results from Bing over multiple scrapes during a period of a few months. In total the final evaluation set contains 171,302 unique documents across all queries which are then judged by human evaluators on a five point relevance scale.

We report the normalized discounted cumulative gain (NDCG) at different rank positions as a measure of performance for the different models. The results show that the $DESM_{IN-OUT}$ outperforms both the BM25 and the LSA baselines, as well as the $DESM_{IN-IN}$ at all rank positions. The embeddings trained on the query corpus also achieves better results than the embeddings trained on body text. We provide additional analysis and experiment results in [4].

4. DISCUSSION AND CONCLUSION

We formulated a *Dual Embedding Space Model* (DESM) that leverages the often discarded output embeddings learned by the word2vec model. Our model exploits both the input and the output embeddings to capture topic-based semantic relationships. The examples in Table1 show that different nearest neighbours can be found by using proximity in the IN-OUT vs the IN-IN spaces. In our experiments ranking via proximity in the IN-OUT space performs better for retrieval than the IN-IN based ranking. This finding emphasizes that the performance of the word2vec model is application dependent and that quantifying semantic relatedness via cosine similarity in the IN space should not be a default practice.

References

[1] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*, pages 3111–3119, 2013.

[4] B. Mitra, E. Nalisnick, N. Craswell, and R. Caruana. A dual embedding space model for document ranking. *arXiv:1602.01137*, 2016.

[5] S. Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5): 503–520, 2004.

[6] S. Robertson and H. Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.