Table 9: NTCIR-12 Results after Combining Approach0 and Tangent-CFT (TanApp) vs. Individual Retrieval Models.

Retrieval result	Partial bpref	Full bpref	Harm. Mean bpref
TanApp	**0.73**	**0.70**	**0.71**
Tangent-CFT	0.71	0.60	0.65
Approach0	0.59	0.67	0.63
Tangent-s	0.59	0.64	0.61

4.5 Combining Tangent-CFT with Approach0

Relying only on embeddings for formula retrieval may lead to higher partial relevance scores, but perhaps at the cost of lower full relevance scores when compared to operating directly on a formula tree. For example, Approach0 retrieves formulas using leaf-root paths in OPTs, and obtains higher full relevance scores than Tangent-CFT (see Table 2).

To try and leverage the strengths of both the embedding-based and tree-based approaches, we created another model (*TanApp*) that linearly combines retrieval scores from Tangent-CFT and Approach0. Given weight parameter $\alpha \in [0, 1]$, TanApp calculates the score for a given query 'q' as follows:

$$Score_q(f) = \alpha \cdot Tangent\text{-}CFT_q(f) + (1 - \alpha) \cdot Approach0_q(f)$$

We used a grid search over alpha to find the weight that maximizes one of three possible target measures: (1) the average partial bpref, (2) the average full bpref, or (3) the average harmonic mean of full and partial bpref using leave-one-out cross-validation. The first row of Table 9 shows the TanApp results optimized to each measure (i.e., three separate grid-searches optimized for partial, full, and harmonic mean of bpref), along with the corresponding results for each individual system. For comparison, Tangent-s results are also shown. As can be seen, TanApp is the best choice, regardless of the chosen evaluation measure.

5 CONCLUSION

In this paper, we presented Tangent-CFT, an embedding model for mathematical formulas. We use fastText to produce formula embeddings for both symbol layout trees (SLTs) that capture formula structure, and operator trees (OPTs) that capture formula semantics. The embedding procedure converts a tree-based formula representation into a sequence of tuples. Each tuple is treated as a word, with its tokenized elements treated as characters. Tuple 'words' are then embedded using n-grams of varying lengths computed over the tuple and its neighboring tuples in the sequence.

Our Tangent-CFT model combines OPT, SLT, and SLT-Type embeddings to obtain higher partial relevance than state-of-the-art models for the NTCIR-12 formula browsing task. We have also shown that combining results from an embedding model such as Tangent-CFT with results from a structure matching approach (e.g., Approach0) can produce higher partial *and* full relevance scores than previous approaches.

For future work, we plan to extend the existing test collection to include more diverse query formulas, and particularly formulas that are not present as exact matches in the collection. Also, we plan to incorporate text near formulas into our embedding model. So far we have only studied isolated formula retrieval, and we expect that leveraging nearby text would further improve the representation, as was observed in the NTCIR MathIR task [19].

Acknowledgements. This material is based upon work supported by the Alfred P. Sloan Foundation under Grant No. G-2017-9827 and the National Science Foundation (USA) under Grant No. IIS-1717997.

REFERENCES

[1] Akiko Aizawa, Michael Kohlhase, Iadh Ounis, and Moritz Schubotz. 2014. NTCIR-11 Math-2 Task Overview. In *In Proceedings of the 11th NTCIR Conference.*
[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics.*
[3] Chris Buckley and Ellen M Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*
[4] Kenny Davila and Richard Zanibbi. 2017. Layout and semantics: Combining representations for mathematical formula search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval.*
[5] Dallas Fraser, Andrew Kane, and Frank Wm Tompa. 2018. Choosing Math Features for BM25 Ranking with Tangent-L. In *Proceedings of the ACM Symposium on Document Engineering 2018.*
[6] Liangcai Gao, Zhuoren Jiang, Yue Yin, Ke Yuan, Zuoyu Yan, and Zhi Tang. 2017. Preliminary Exploration of Formula Embedding for Mathematical Information Retrieval: can mathematical formulae be embedded like a natural language?
[7] Giovanni Yoko Kristianto, Goran Topic, and Akiko Aizawa. 2016. MCAT Math Retrieval System for NTCIR-12 MathIR Task. In *NTCIR.*
[8] Kriste Krstovski and David M Blei. 2018. Equation Embeddings.
[9] P Pavan Kumar, Arun Agarwal, and Chakravarthy Bhagvati. 2012. A structure based approach for mathematical expression retrieval. In *International Workshop on Multi-disciplinary Trends in Artificial Intelligence.*
[10] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning.*
[11] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research.*
[12] Behrooz Mansouri, Douglas W. Oard, and Richard Zanibbi. 2019. Characterizing Searches for Mathematical Concepts. In *Joint Conference on Digital Libraries.*
[13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.
[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems.*
[15] Bhaskar Mitra and Nick Craswell. 2015. Query auto-completion for rare prefixes. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management.*
[16] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM.
[17] Petr Sojka and Martin Líška. 2011. The art of mathematics retrieval. In *Proceedings of the 11th ACM Symposium on Document Engineering.*
[18] Abhinav Thanda, Ankit Agarwal, Kushal Singla, Aditya Prakash, and Abhishek Gupta. 2016. A Document Retrieval System for Math Queries. In *NTCIR.*
[19] Richard Zanibbi, Akiko Aizawa, Michael Kohlhase, Iadh Ounis, Goran Topic, and Kenny Davila. 2016. NTCIR-12 MathIR Task Overview. In *NTCIR.*
[20] Richard Zanibbi and Dorothea Blostein. 2012. Recognition and retrieval of mathematical expressions. *International Journal on Document Analysis and Recognition (IJDAR).*
[21] Richard Zanibbi, Kenny Davila, Andrew Kane, and Frank Wm Tompa. 2016. Multi-stage math formula search: Using appearance-based similarity metrics at scale. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval.*
[22] Wei Zhong and Richard Zanibbi. 2019. Structural Similarity Search for Formulas Using Leaf-Root Paths in Operator Subtrees. In *European Conference on Information Retrieval.*