Table 3: Online A/B test metrics on different groups of job seekers. Lift is w.r.t non-personalized query suggestions.

Metric	Job seeker group	Lift (p value)
search sessions	passive job seeker	+1.19% (2.1×10^{-3})
	active job seeker	-0.49% (0.1)
successful search sessions	passive job seeker	+1.24% (0.02)
	active job seeker	-0.52% (0.14)

CTR-r also improved **0.3%** (p-value 0.02) compared with the non-personalized Seq2Seq model. This result shows that personalized query suggestions are more engaging and improve the overall user search experience.

At LinkedIn, we define a search session by the start of a new search query on the homepage or a time gap of user-inactivity. A successful session is one in which the user performed some meaningful action, such as saving a job or applying for a job. The number of successful search sessions is one fundamental business metric for LinkedIn. We classify users into active and passive job seekers based on their recent job-related activities, such as applying for a job. Moreover, we are interested in how changes in the search engine impact job seeker behaviors. Table 3 shows how personalization impacts different groups of job seekers.

Notably, personalization has a more significant impact on passive job seekers. We believe passive job seekers are more easily encouraged to follow the personalized query suggestions to start new searches because those suggestions are more relevant to their background and intent. For the same reason, those query suggestions can retrieve more attractive search results, leading to a better overall search experience. Active job seekers, on the other hand, may already have a good idea for what they want. The query suggestions provided by our model, therefore, has less impact on them.

5 ONLINE SERVING STRATEGY

A significant challenge for applying a DNN to production online inference is the large latency it incurs. However, in our particular application, the Q.S. service and the search retrieval and ranking process can be parallelized because they are independent of each other. We also added an in-memory cache to help reduce latency because the same input source query should always output the same target query suggestions for the same model. Figure 4 shows our online serving architecture. This strategy worked well: With tens of millions of requests served per day, we achieved an average latency of 23 ms and a 99^{th} percentile of 70 ms.

6 CONCLUSIONS

In this work, we presented a query suggestion framework based on recurrent neural networks. The framework models both structured user features as well as unstructured text data. In our offline experiments, we showed that modeling both data gave a better performance than modeling text data alone. We further showed that treating the categorical user-feature as an "additional vocabulary" not only was straightforward to implement but also gave better performance compared with concatenating the user feature embedding to word embeddings.

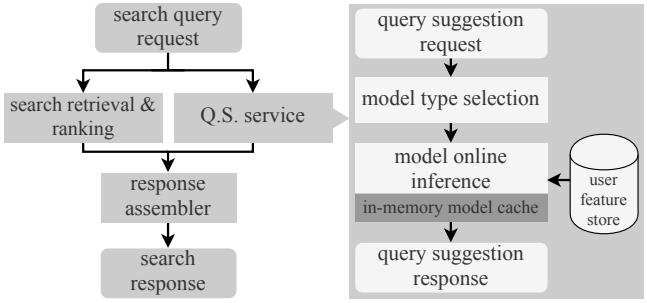

Figure 4: Online serving architecture of Q.S. service.

A/B test on live traffic confirmed our offline experiment observations. Personalization boosted core business metrics for LinkedIn search. Different user segments benefited differently from personalized query suggestions. We showed that passive job seekers are more likely to benefit from better query suggestions experience.

This framework is a production-grade, deep-learning–based query suggestion framework. The online serving strategy presented in this work enabled us to make real-time inferences from deep learning models. Our system is general and can be applied to other search engines as well.

REFERENCES

[1] Hiteshwar Kumar Azad and Akshay Deepak. 2017. Query expansion techniques for information retrieval: A survey. *Information Processing & Management* 56, 5 (Aug 2017), 1698–1735.

[2] Mohamed Reda Bouadjenek, Hakim Hacid, Mokrane Bouzeghoub, and Johann Daigremont. 2011. Personalized social query expansion using social bookmarking systems. In *SIGIR '11*. ACM Press, New York, New York, USA, 1113.

[3] Heng-Tze Cheng, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, Hemal Shah, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, and Wei Chai. 2016. Wide & Deep Learning for Recommender Systems. In *DLRS 2016*. ACM Press, New York, New York, USA, 7–10.

[4] Paul Alexandru Chirita, Claudiu S. Firan, and Wolfgang Nejdl. 2007. Personalized query expansion for the web. In *SIGIR '07*. New York, New York, USA.

[5] Donna Harman. 1988. Towards interactive query expansion. In *SIGIR '88*. ACM Press, New York, New York, USA, 321–331.

[6] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* (1997).

[7] Aaron Jaech and Mari Ostendorf. 2018. Low-Rank RNN Adaptation for Context-Aware Language Modeling. *Transactions of the Association for Computational Linguistics* 6 (Dec 2018), 497–510.

[8] J. Jayanthi, K. S. Jayakumar, and B. Akalya. 2011. Personalized Query Expansion based on phrases semantic similarity. In *ICECT '11*. IEEE, 273–277.

[9] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics* 5 (Dec 2017), 339–351.

[10] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. *EMNLP '15* (Aug 2015), 1412–1421.

[11] Qiaozhu Mei, Dengyong Zhou, and Kenneth Church. 2008. Query suggestion using hitting time. In *CIKM '08*. ACM Press, New York, New York, USA, 469.

[12] Tomas Mikolov and Geoffrey Zweig. 2012. Context dependent recurrent neural network language model. In *SLT '12*. IEEE, 234–239.

[13] Mandar Mitra, Amit Singhal, and Chris Buckley. 1998. Improving automatic query expansion. In *SIGIR '98*. ACM Press, New York, New York, USA, 206–214.

[14] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob G. Simonsen, and Jian-Yun Nie. 2015. A Hierarchical Recurrent Encoder-Decoder For Generative Context-Aware Query Suggestion. *CIKM '15* (Jul 2015), 553–562.

[15] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. *NIPS '14* (Sep 2014).

[16] Zhengyu Zhu, Jingqiu Xu, Xiang Ren, Yunyan Tian, and Lipei Li. 2007. Query Expansion Based on a Personalized Web Search Model. In *SKG 2007*. IEEE, 128–133.