Multi-lingual Semantic Search for Domain-specific Applications: Adobe Photoshop and Illustrator Help Search

Jayant Kumar
jaykumar@adobe.com
Adobe Inc.
USA

Ashok Gupta
ashogupt@adobe.com
Adobe Inc.
USA

Zhaoyu Lu
lolu@adobe.com
Adobe Inc.
USA

Andrei Stefan
astefan@adobe.com
Adobe Inc.
USA

Tracy Holloway King
tking@adobe.com
Adobe Inc.
USA

ABSTRACT

Search has become an integral part of Adobe products and users rely on it to learn about tool usage, shortcuts, quick links, and ways to add creative effects and to find assets such as backgrounds, templates, and fonts. Within applications such as Photoshop and Illustrator, users express domain-specific search intents via short text queries. In this work, we leverage sentence-BERT models fine-tuned on Adobe's HelpX data to perform multi-lingual semantic search on help and tutorial documents. We used behavioral data (queries, clicks, and impressions) and additional annotated data to train several BERT-based models for scoring query-document pairs for semantic similarity. We benchmarked the keyword-based production system against semantic search. Subsequent AB tests demonstrate that this approach improves engagement for longer queries while reducing null results significantly.

CCS CONCEPTS

• Information systems → Retrieval models and ranking; **Content analysis and feature selection**; **• Applied computing → Document management and text processing**.

KEYWORDS

semantic search, fine-tuning text embeddings, null result recovery

ACM Reference Format:
Jayant Kumar, Ashok Gupta, Zhaoyu Lu, Andrei Stefan, and Tracy Holloway King. 2023. Multi-lingual Semantic Search for Domain-specific Applications: Adobe Photoshop and Illustrator Help Search. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3591826

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591826

1 CUSTOMER PROBLEM

Search and discovery[1] have become an integral part of creative product experiences, helping customers to learn how to use products and tools and to find assets such as Stock images, templates, and fonts. Users can search for product help and tutorials using the Adobe Help Center, known as "HelpX" [7], which is powered by Adobe's Universal Search Service (USS). Since there are only ~25K HelpX documents across 21 languages and since the domain is highly specific (e.g. how to use Adobe's editing tools), matching the user's query to relevant documents is challenging but essential for users to be able to use Adobe's products.

Traditional search systems using keywords and BM25-based rankers perform well for shorter queries. However, they do not perform as well for longer queries because they treat the query as a bag of words. For example, consider the two queries in (1):

(1) a. white rose on red background
  b. red rose on white background

Keyword-based search systems typically return similar results for both queries, potentially with different rankings due to token proximity features. Determining the intent behind long queries requires a different approach using embeddings, referred to as semantic search [6].

Users have become used to powerful search engines like Google and Bing that can handle complex natural language queries. As a result, in applications like Adobe products, users have started using longer, multi-word queries to express their intent. Our work on semantic search enables Adobe's help and tutorial search system to process longer, multi-word queries in different languages to provide a seamless, highly relevant user experience. Our first application of semantic search was for help and hands-on tutorials in the Photoshop Discover Panel for English queries.[2] To evaluate its effectiveness, we ran an AB test that compared results from the new semantic search system with the existing keyword-based system. The AB test showed an increase in click-through rate (CTR) and a reduction in null rate for longer queries. We then adjusted the

[1]We would like to thank the broader Adobe search and discovery team as well as the Adobe in-app experience team for their support in bringing semantic search to production. We especially thank Kerem Turgutlu, who worked closely on the multi-lingual aspects of this project while at Adobe.

[2]The Discover Panel is a component of Adobe applications which allows users to search for help and assets while still using the creative apps. The results are contextualized to the app and personalized to the user. The UX for the Discover Panel is shown in Figure 1, with the panel separate from the in-app editing pane.

definition of long queries to include more queries and extended the search to three more languages (French, German, and Japanese) and then to a further 17 languages. Finally, we extended semantic search to Adobe Illustrator, which has a different set of help documents but a similar vocabulary.

We trained the semantic search model using a sentence-BERT Siamese network [11] (section 3) fine-tuned for the Adobe help in-domain vocabulary. The fine-tuning dataset included past query-document click pairs as well as similar documents detected using an off-the-shelf BERT model (section 2). Internal human evaluation was used to select which queries to apply semantic search to, to determine which model performed best and to eliminate false-positives in the similar document data.

In this paper we report on how we built on recent advances in NLP language modeling and semantic search by:

- Fine-tuning the model for a highly specialized domain, thereby confirming academic research on the value of fine-tuning
- Building a multi-lingual model, thereby confirming academic research on how multi-lingual models can provide higher accuracy even in mono-lingual uses
- Leveraging past user behavioral data, machine translation, and off-the-shelf models to create custom training data
- Using human annotation to determine which query classes to apply semantic search to
- Building a simple but effective low-latency hybrid semantic and keyword search system
- Iteratively AB testing in order to safely but quickly improve the user experience

2 DATA PREPARATION

In order to train a model that works effectively for the vocabulary found in the help and tutorial documentation and to map it to the terminology used in queries, we needed pairs of text samples with similarity scores. We collected data from two sources: past user behavioral data and semi-automatic document similarity data.

We mined HelpX queries and the corresponding clicks on tutorial documents from a one year period. The data was converted to <query, title, score> triplets for training. This gave us ~12K triplets for English (see section 5 for extensions beyond English).

We also computed the top nearest-neighbor of titles for all tutorial documents using the open-source sentence-BERT model [11]. We used this data to create triplets <title1, title2, score> of semantically similar in-domain documents. We manually annotated the resulting triplets to fine-tune the sentence-BERT model, resulting in ~8K triplets. In total, we used ~20K triplets for fine-tuning.

3 MODEL TRAINING AND OFFLINE EVALUATION

The model is built on BERT (Bidirectional Encoder Representations from Transformers) language models [4, 5]. We used an open-source pre-trained BERT model and fine-tuned it with Adobe-specific content from HelpX tutorials [7] (section 2). We tried several base models including RoBERTa-large [9] and DistilBERT-base [12]. The best model was selected based on an evaluation set collected during an internal evaluation. Due to the domain-specific knowledge in

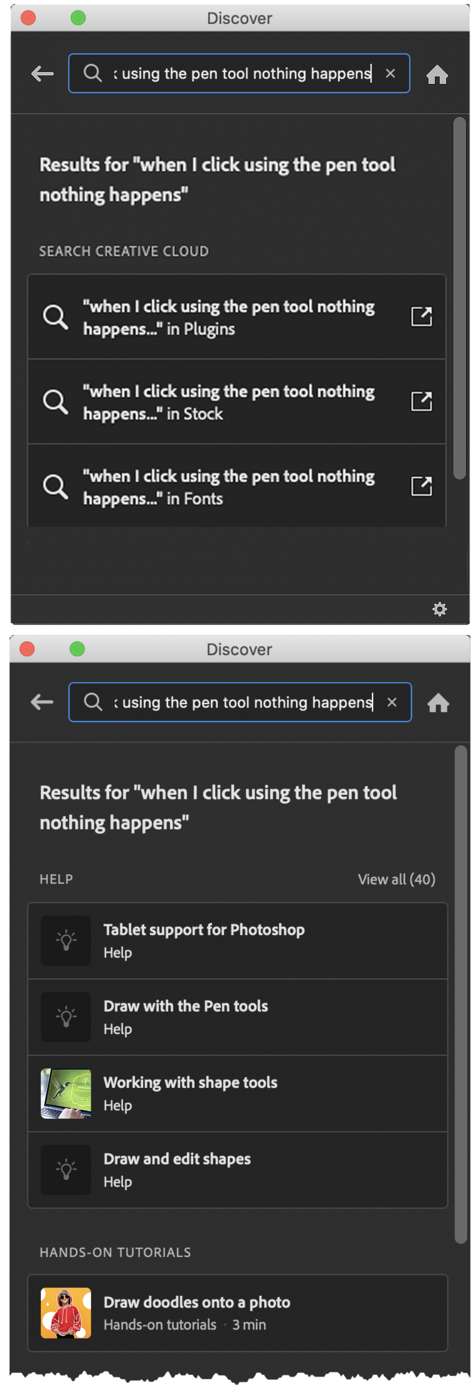

Figure 1: Results for *when i click using the pen tool nothing happens* in the Photoshop Discover Panel before (top) where there are no results and the user is prompted to search other Adobe surfaces and after (bottom) using semantic, embedding-based search with help and tutorial results.

the help and tutorials for Adobe applications, it was not possible to use public crowds to annotate data.

The model semantically encodes queries and documents as embeddings. For the documents, we tried several variants for encoding: the title and the short description of the documents individually and in concatenated form and encoding the whole document via paragraph encoding. Titles are generally 5–20 words long, while the descriptions are generally 20–50 words long and whole documents can be multiple pages long. An example is shown in (2).

(2) **Title:** Increase midtone contrast with Curves
Description: Learn how to use Curves adjustments in Photoshop to get advanced control over midtones. Also explore Curves presets and the basics of blend modes.

The best performance was obtained using a combination of the title and description of the document.

Text queries are similarly embedded. The model then optimizes the mean-squared loss objective function by bringing the embeddings of similar queries and documents together while pushing the dissimilar ones apart in the learned embedding space (Figure 2). To return search results for a given query, we compute a similarity score between the query and document embeddings. This score is used to rank the documents with minor adjustments based on recency (i.e. date of document creation) and behavioral data features.

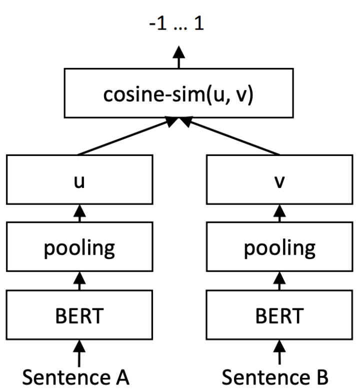

Figure 2: SBERT training architecture. Two sentences in a training batch are passed through the network and the cosine similarities are computed. The mean squared loss is minimized based on the sentence similarity scores. The two BERT networks have tied weights (Siamese network structure).

Human evaluation was used to compare the models offline to the production results and to each other. All evaluation was done internally since domain expertise was required to judge the relevance of results. A standard 5-point relevance scale was used. NDCG was computed for the top 4 results (NDCG@4) for 150–175 queries. The queries were sampled from user query logs along dimensions of: length, head/torso/tail frequency, and null queries. Although the query sample size was small, this allowed us to rapidly determine that pure semantic search was not effective for short queries (section 4) and that embeddings of the title and short description (see (2)) were more effective than ones including all the document text.

4 INTEGRATION

The semantic search component was integrated into the overall search system. One option was to replace the existing keyword-based help search with semantic search. However, human evaluation of a stratified sample of queries rapidly showed that short queries (1–2 words) had worse results when using semantic search. This reflects the fact that these short queries (e.g. *remove background*, *resize*) are relatively frequent and so can take advantage of user behavioral data to determine the best results [1]. In addition, these short queries did not contain enough text to differentiate among the multiple tutorials that covered similar topics. So, based on the overall relevance and null rate for different query lengths, we created a hybrid system whereby queries >=3 words were sent to the semantic search system for recall and ranking, while short queries were sent to the existing keyword-based search (Figure 3).

The semantic search system has three components:

- Query service: Adobe's Universal Search Service (USS), scoped to tutorials for the application (Photoshop or Illustrator)
- A cloud-based service that computes semantic embeddings using the fine-tuned BERT model
- An Elasticsearch index of keywords, embeddings and metadata for the help and tutorials

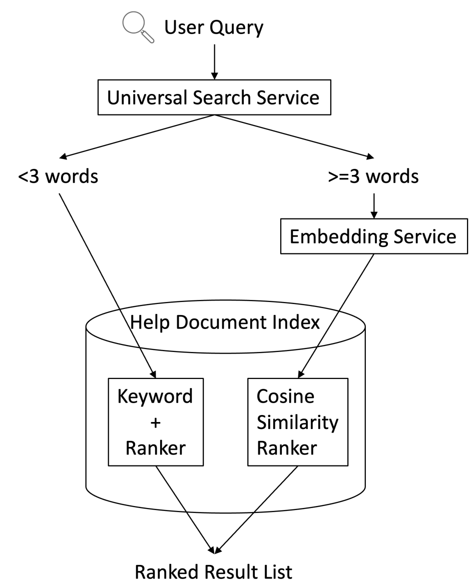

Figure 3: Integration of semantic search with traditional keyword search (section 4).

The cloud-based service for computing semantic embeddings is used in batch mode at index time to compute the embeddings for the documents and store them for comparison to the query embeddings. It is also used at query time to compute the embedding for the query on the fly: since many of these queries (e.g. *how to create a heart shape*, *feather the edges of a photo*) are unique within a reasonable timeframe (e.g. a week), they cannot be precomputed. However, the fact that even longer queries are relatively short means that latencies are low (see [2] on the importance of efficiency in production IR systems). In addition, there are only $\sim$25K documents and hence no optimization had to be done to meet the low latency requirements.

5 MULTI-LINGUAL EXTENSION

After the success of the English AB test (section 6), we leveraged the click data of query-document pairs from three additional languages: French, German, and Japanese. The query-related data, especially the query-document click data, was much sparser for languages other than English. We used cross-language query translation to increase the training dataset size by a factor of 12, which improved retrieval performance among all the languages, especially for the lower resourced ones. Since the queries are >=3 words, the quality of the machine translation was sufficient for training despite the domain-specific nature of the queries.

A single model is used for all of the languages and language is not used directly as a feature during training. The multi-lingual model was trained by fine-tuning the pretrained XLM-RoBERTa encoder [3] with InfoNCE contrastive loss [10]. We ran multiple experiments and evaluations to find the optimal model to attain high retrieval performance on queries from the four languages (English, French, German, Japanese). As with English, the models were compared with NDCG@4 since only a few results are shown to the user and hence the top results are the most important (Figure 1). Since only longer queries are sent to the semantic search component, no language detection [8, 13] had to be applied to the query: the multiple query words are distinct enough to map correctly into the document embedding space regardless of language. The final system displays the document in the user's selected language (e.g. the app-specified language) regardless of the query language.

Once these four languages were in production, we trained a model for 17 more languages (21 languages total). Scaling manual evaluation across 21 languages was prohibitive. For the first language expansion round to French, German, and Japanese, we replicated the manual evaluation used for English (section 3). However, for the remaining 17 languages, we adjusted the offline evaluation to ensure that we would do no harm in launching semantic search. To do this, we looked at the overlap between results from production keyword search and from semantic search and at the null rate reduction. We also used machine translation of queries into English to evaluate the other languages.

The evaluations indicated that having more data from different languages helps the model understand semantic associations better than a mono-lingual model. As a result, in addition to using the multi-lingual model for non-English languages, we replaced the original mono-lingual English model with the new, multi-lingual one.

6 AB TEST RESULTS

Our first application of semantic search was for help and tutorials in the Photoshop Discover Panel for English only. To evaluate its effectiveness, we ran an AB test that compared results from the semantic search system with the existing keyword-based system. Since the initial AB test only affected queries with >=4 words, we focused on the click-through rate (CTR) and null rate for these queries. CTR was measured simply as clicks per issued query. Null rate was measured as the number of search result pages with no results returned per the number of search result pages (i.e. the number of queries). The null rate was not reduced to 0 since a similarity threshold was applied to the semantic search: only results with a sufficiently high score were returned. This prevented the system from returning irrelevant results when there was no relevant document for the user's query.

Photoshop

Metric	English		Fr/De/Jp	+17 lang
	>=4 words	>=3 words	(>=3 words)	
CTR	+5.2	+3.0	+8.0	+8.6
Null Rate	−4.5	−6.0	−16.6	−30.9

Illustrator

Metric	All 21 languages
CTR	+9.1
Null Rate	−21.0

Table 1: Results of AB test on Photoshop (top) and Illustrator (bottom) Discover Panels. Changes are in percentage points. Change in CTR is for treated queries (>=3 or 4 words) only.

The AB test showed a statistically significant improvement in CTR and null rate for long queries. In addition, we looked at the overall CTR to confirm that we had not harmed the shorter queries (e.g. through introducing a bug): as expected, the shorter, more frequent queries were unaffected by the integration of the semantic search component. We then AB tested applying semantic search to queries with >=3 words. Having determined the best model and hybrid system for English, we AB tested language expansion to French, German, and Japanese. Once those proved successful, we tested the next 17 languages. Based on the results for Photoshop, we tested all 21 languages for queries >=3 words for Adobe Illustrator, a similar tool used for creating illustrations. The results are shown in Table 1.[3]

7 CONCLUSION AND NEXT STEPS

This paper looked at semantic search for Adobe Photoshop and Illustrator help queries. The system takes advantage of recent advances in language modeling and semantic search as well as classic keyword-based search by applying semantic search only to longer queries and leveraging user behavioral data and machine translation to create large scale training data.

We are continuing our work by adding more types of help and tutorial content to the result sets for Photoshop and Illustrator and by expanding to surfaces with broader types of help content such as overall Creative Cloud and Adobe.com search.

ADOBE COMPANY PORTRAIT

Adobe Inc. enables customers to change the world through digital experiences and creativity. The Adobe search and discovery team supports search and recommendations across customer text, image,

[3]We are unable to publish the exact CTR and null rate. We show the results in Table 1 as percentage point change. For example, a +5 in CTR indicates going from a CTR of X% (e.g. 60%) to a CTR of X+5% (e.g. 65%).

video, and other document types, as well as over Adobe Stock assets and Adobe help and tutorials.

MAIN AUTHOR BIO

Jayant Kumar is a senior staff applied-ML scientist at Adobe Inc. He earned his Ph.D. from University of Maryland, College Park in 2013. He has over 25 publications in peer-reviewed conferences and journals. He has served as a PC member for Computer Vision conferences including CVPR, ICCV and ICML. He holds 20 patents.

REFERENCES

[1] Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 19–26. https://doi.org/10.1145/1148170.1148177

[2] Sebastian Bruch, Claudio Lucchese, and Franco Maria Nardini. 2022. ReNeuIR: Reaching Efficiency in Neural Information Retrieval. In *SIGIR '22: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 3462–3465. https://doi.org/10.1145/3477495.3531704

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[4] Jacob Devlin and Ming-Wei Chang. 2018. Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html Google AI Blog.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. ACL, 4171–4186. arXiv:1810.04805v2.

[6] R. Guha, Rob McCool, and Eric Miller. 2003. Semantic Search. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*. ACM, 700–709.

[7] Adobe Inc. 2023. HelpX. https://helpx.adobe.com/support.html Adobe online public help pages.

[8] Tracy Holloway King, Chirag Arora, Francois Guerin, Sachin Kelkar, and Judy Massuda. 2021. The Last Mile: Taking Query Language Identification from Model Ready to Production. In *SIGIR ECOM21: Workshop on Ecommerce Search*. CEUR.

[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://doi.org/10.48550/ARXIV.1907.11692

[10] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. https://doi.org/10.48550/ARXIV.1807.03748

[11] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. ACL, 3982–3992. arXiv preprint arXiv:1908.10084.

[12] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. https://doi.org/10.48550/ARXIV.1910.01108

[13] Ritiz Tambi, Ajinkya Kale, and Tracy Holloway King. 2020. Search Query Language Identification Using Weak Labeling. In *Proceedings of LREC 2020*. LREC, 3513–3520.