Turiya at PerpectiveArg2024: A Multilingual Argument Retriever and Reranker

Sougata Saha and Rohini Srihari
State University of New York at Buffalo
Department of Computer Science and Engineering
{sougatas, rohini}@buffalo.edu

Abstract

While general argument retrieval systems have significantly matured, multilingual argument retrieval in a socio-cultural setting is an overlooked problem. Advancements in such systems are imperative to enhance the inclusivity of society. The Perspective Argument Retrieval (PAR) task addresses these aspects and acknowledges their potential latent influence on argumentation. Here, we present a multilingual retrieval system for PAR that accounts for societal diversity during retrieval. Our approach couples a retriever and a re-ranker and spans multiple languages, thus factoring in diverse socio-cultural settings. The performance of our end-to-end system on three distinct test sets testify to its robustness.

1 Introduction

Given a query, argument retrieval (Manning, 2008; Bondarenko et al., 2020, 2022) generally involves retrieving a set of k-relevant arguments from a corpus. Perspective argument retrieval (PAR) (Falk et al., 2024) is an expansion and concerns factoring in the socio-cultural factors during retrieval. Apart from the semantic features, it considers aspects such as persona, attitude, demographics, etc, during retrieval. Such systems are imperative to expanding the reach of argumentation technologies (Besnard and Hunter, 2008; Van Eemeren et al., 2015) among diverse socio-cultural groups. We tackle the following two scenarios of the PAR shared task: (i) **Baseline:** Given a query, retrieving the relevant arguments from a corpus. This scenario evaluates the general abilities of a system to retrieve relevant arguments. (ii) **Explicit:** This extends the baseline task by explicitly adding socio-cultural information to the query and the corpus and limiting the relevant candidates to arguments from authors matching the corresponding socio-cultural background. This scenario tests if a retrieval system can consider socio-cultural properties when explicitly mentioned in the query and the candidates.

The argument corpus comprises 26,335 arguments covering the 2019 Swiss Federal elections in German, French, and Italian. Each argument is enriched with eight socio-cultural properties and spans 45 political aspects. The queries are political issues and based on the x-stance dataset (Vamvas and Sennrich, 2020). The training queries span 35 political aspects, whereas the development set queries span the other 10 aspects. The final evaluation set comprises three secret test sets.

2 Proposed Method

Our implemented architecture comprises a retriever and a re-ranker. Figure 1 illustrates our architecture, which we explain in detail below.

2.1 Corpus Processing

We use *Mistral-7B-Instruct-v0.2* (Jiang et al., 2023) (henceforth referred to as Mistral) in a zero-shot setting to first translate all arguments in the corpus to English. The zero-shot prompt is *"Translate the following text to English."*. Next, we use the *multi-qa-mpnet-base-dot-v1* (Reimers and Gurevych, 2019) model to generate the English embeddings and the *paraphrase-multilingual-mpnet-base-v2* model for multilingual embeddings for each argument. We populate a graph-based vector index (Hsnwlib[1] (Malkov and Yashunin, 2018)) with the English embeddings for performing an approximate K-nearest neighbor search during retrieval. The Hsnwlib index and the multilingual embeddings comprise our retrieval argument collection.

2.2 Retriever

We represent a query using two embeddings: (i) We translate the multilingual query to English using Mistral and generate its English-translated

[1] https://github.com/nmslib/hnswlib

Proceedings of the 11th Workshop on Argument Mining, pages 159–163

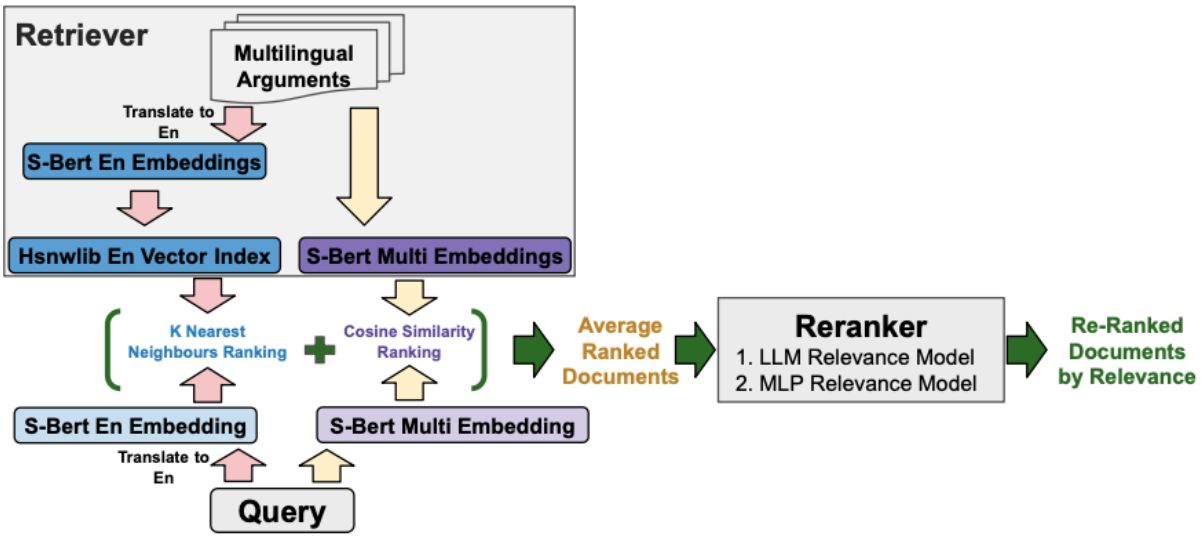

Figure 1: End-to-end architecture of the retrieval system.

embedding using the *multi-qa-mpnet-base-dot-v1* model. (ii) We generate the multilingual embedding from the original non-translated query using the *paraphrase-multilingual-mpnet-base-v2* model. Cosine similarity is computed between the query and corpus multilingual embeddings to retrieve the top 1000 most similar arguments. K-nearest neighbor (KNN) search is performed on the Hsnwlib index to retrieve the top 1000 similar arguments to the English-translated embedding. We average the two rankings and order them to yield the final top 1000 documents. The choice of averaging the multilingual embedding-based cosine similarity ranking with the translated KNN-based ranking is determined empirically. Listed in Table 1, we experimented with different combinations and chose the combination that yielded the best NDCG (Wang et al., 2013) (Normalized discounted cumulative gain) score on the development set.

Id	Combination	4	8	16	20
1	cos_multi	0.97	0.96	0.96	0.95
2	cos_en	0.98	0.96	0.95	0.95
3	knn_multi	0.94	0.93	0.91	0.90
4	knn_en	0.97	0.97	0.97	0.96
5	cos_en + cos_multi	0.97	0.98	0.98	0.98
6	knn_en + knn_multi	**1.00**	**0.99**	0.97	0.97
7	cos_multi + knn_multi	0.96	0.96	0.96	0.95
8	cos_en + knn_en	0.98	0.97	0.96	0.95
9	cos_multi + knn_en	**1.00**	**0.99**	**0.99**	**0.99**
10	cos_en + knn_multi	0.96	0.97	0.96	0.96
11	cos_en + cos_multi + knn_en	0.99	0.98	0.98	0.98
12	cos_en + cos_multi + knn_multi	0.96	0.97	0.97	0.96
13	cos_en + knn_en + knn_multi	0.99	**0.99**	0.97	0.97
14	cos_multi + knn_en + knn_multi	0.99	**0.99**	0.98	0.98
15	cos_en + cos_multi + knn_en + knn_multi	0.98	0.98	0.98	0.98

Table 1: Dev set NDCG results of different scoring combinations for the baseline scenario. Best scores highlighted in bold. cos: cosine similarity, knn: K-nearest neighbours, multi: multilingual, en: English.

2.3 Re-ranker

The re-ranker is an ensemble of a Large Language Model (LLM (Chang et al., 2023; Min et al., 2023; Hadi et al., 2023))-based and a Multi-Layered Perceptron (MLP)-based relevance model. Given a list of the 1000 retrieved documents to a query, we re-order the top 30 by persisting the ones deemed relevant to the query by both models. Below, we detail each model and our internally constructed dataset to train them.

2.3.1 Dataset Construction

We run the retriever on the training set and bucket the retrieved documents by their ranking as follows: 1-20, 21-100, 101-300, 301-700, 700-1000. For each bucket, we check if the retrieved documents are relevant and construct a balanced dataset (named *rel*) of 6,150 examples (4,815 train, 1,335 dev) comprising query and document pairs with a binary label denoting whether the document is relevant to the query. We consider the English-translated text of the query and document. Furthermore, we construct 8,795 examples (6,895 train, 1,900 dev, and named *mcq*) where, given a query

and two retrieved documents, the task requires comparing the documents to determine the more relevant one. From each bucket, we randomly pair relevant and non-relevant retrieved documents. Additionally, we include two cases where random pairs of documents from the same bucket within a window of 3 are marked relevant or non-relevant to the same degree. Constructing the *rel* and *mcq* datasets using the bucketed approach helps adjust the dataset's difficulty, where examples from the lower buckets are more challenging than the higher ones.

2.3.2 MLP Re-ranker

The MLP-based relevance model is a 2-layered neural network. It inputs 768-dimensional query and document embeddings and independently encodes them to a 128-dimensional representation using a single-layered neural network, followed by a non-linear ReLU activation. The encoded representations are concatenated (now 256-dimensional) and passed through a 2-layered neural network, where the hidden layer contains 128 nodes with ReLU activation, and the final layer is a single node that denotes the relevance score (logit). We use *multi-qa-mpnet-base-dot-v1* to compute the 768-dimensional input query and document embeddings. The model is trained on the *rel* dataset in mini-batches of 32 with AdamW (Loshchilov and Hutter, 2017) optimizer, using a 1e-5 learning rate and early stopping for five epochs. It attains an F1 score of 73% on the *rel* dataset dev split.

2.3.3 LLM Re-ranker

We fine-tune Mistral-7B-Instruct-v0.2 on instructions from the *rel* and *mcq* datasets. Table 2 illustrates samples from each dataset. The model was trained for two epochs using LoRA (Hu et al., 2021), a parameter-efficient fine-tuning (Mangrulkar et al., 2022) method. The LoRA r and alpha were set to 16 and 32 and trained the q, v, k, o, gate, up, and down projection modules of the attention heads and the LM head using a 2.5e-5 learning rate. Fine-tuning LLMs on multiple tasks has shown to be fruitful. Since the *rel* dataset instructions only entail comparing query-document pairs, we include the *mcq* dataset to increase the task variety as it additionally entails comparing documents. Fine-tuning on multiple tasks (multi-task learning (Caruana, 1997; Zhang and Yang, 2021)) has proven to improve performance on the individual tasks. The fine-tuned model attains an F1 score of 79% on the *rel* dataset and an accuracy score of 51% on the *mcq* dataset dev splits.

During inference we ensemble the MLP and LLM-based re-rankers, where given a list of the 1000 retrieved documents to a query, we re-order the top 30 by persisting the ones deemed relevant to the query by both models.

Sample from the *rel* dataset
[INST] Identify using true/false if a document is relevant to a query. Query: Should the Confederation support foreigners in integration? Document: "Integration is good and important. However, it is now necessary at the cantonaland community levels." [/INST] Answer: true
Sample from the *mcq* dataset
[INST] Given a query and two retrieved documents, identify which of the following options is correct. Query: Should the Confederation support foreigners in integration? Document A: Political co-determination promotes foreigner integration. Document B: It is desirable to integrate foreigners on a political level. i. Document A is more relevant than Document B. ii. Document B is more relevant than Document A. iii. Both the documents are equally relevant to the query. iv. None of the documents are relevant to the query. [/INST] Answer: iv

Table 2: Samples of Mistral training instructions.

3 Results

We run our pipeline on the three official test sets and share the results for the baseline scenario in Table 3 and the explicit scenario in Table 4. We compare our results against BM-25-based and embedding cosine similarity-based (Sbert) baselines. For the baseline scenario (Table 3), our implementation significantly outperforms both baselines in test sets 2 and 3. Although we significantly outperform the BM-25 baseline for test set 1, the Sbert baseline attains a comparable score to our implementation. For the explicit scenario, we only persist the baseline ranked documents where the query socio-cultural features match with the document. As evident from Table 4, our implementation significantly outperforms the Sbert baseline for all test sets.

4 Conclusion

Here, we present an end-to-end retrieval and ranking system capable of retrieving multilingual arguments to user queries while factoring in the socio-cultural features. Our implementation uses the original and English-translated text and implements an ensembled retriever and re-ranker to retrieve relevant documents. Our retriever combines the semantic relatedness of embedding a similarity-based

		Ndcg				Precision			
Set	Model	4	8	16	20	4	8	16	20
1	BM25	0.72	0.67	0.62	0.60	0.68	0.64	0.58	0.55
	Sbert	**0.99**	**0.99**	**0.98**	**0.98**	**0.99**	**0.99**	**0.98**	**0.98**
	Ours	**0.99**	0.98	**0.98**	0.97	0.98	0.97	0.97	0.97
2	BM25	0.78	0.76	0.72	0.69	0.78	0.76	0.69	0.66
	Sbert	0.88	0.86	0.84	0.84	0.89	0.86	0.82	0.82
	Ours	**0.94**	**0.93**	**0.91**	**0.90**	**0.94**	**0.93**	**0.90**	**0.89**
3	BM25	0.36	0.37	0.38	0.37	0.36	0.37	0.39	0.37
	Sbert	0.67	0.68	0.61	0.59	0.69	0.69	0.59	0.56
	Ours	**0.80**	**0.80**	**0.73**	**0.70**	**0.82**	**0.81**	**0.70**	**0.67**

Table 3: Model performance on diverse test sets for the Baseline scenario.

		Ndcg				Precision			
Set	Model	4	8	16	20	4	8	16	20
1	Sbert	0.22	0.22	0.23	0.23	0.22	0.22	0.22	0.22
	Ours	**0.71**	**0.69**	**0.67**	**0.66**	**0.66**	**0.60**	**0.54**	**0.52**
2	Sbert	0.15	0.15	0.15	0.15	0.15	0.14	0.13	0.13
	Ours	**0.72**	**0.69**	**0.65**	**0.64**	**0.67**	**0.60**	**0.53**	**0.50**
3	Sbert	0.38	0.39	0.42	0.44	0.38	0.36	0.32	0.31
	Ours	**0.70**	**0.65**	**0.62**	**0.61**	**0.67**	**0.58**	**0.49**	**0.47**

Table 4: Model performance on diverse test sets for the Explicit scenario.

approach with a KNN-based approach to yield an initial retrieved ordering of documents. Our ensemble of LLM and MLP-based re-rankers re-orders the documents by their relevance to generate the final list of ordered documents for a query. Evaluations against two baselines across three distinct test sets testify to the robustness of our approach.

References

Philippe Besnard and Anthony Hunter. 2008. *Elements of argumentation*, volume 47. MIT press Cambridge.

Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, et al. 2020. Overview of touché 2020: argument retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11*, pages 384–395. Springer.

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Shahbaz Syed, Timon Gurcke, Meriem Beloucif, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, et al. 2022. Overview of touché 2022: argument retrieval. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 311–336. Springer.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28:41–75.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Neele Falk, Andreas Waldis, and Iryna Gurevych. 2024. Overview of perpectivearg2024: The first shared task on perspective argument retrieval. In *Proceedings of the 11th Workshop on Argument Mining*, Bangkok. Association for Computational Linguistics.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Christopher D Manning. 2008. *Introduction to information retrieval*. Syngress Publishing,.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection. *arXiv preprint arXiv:2003.08385*.

Frans H Van Eemeren, Frans H van Eemeren, Sally Jackson, and Scott Jacobs. 2015. Argumentation. *Reasonableness and effectiveness in argumentative discourse: Fifty contributions to the development of Pragma-dialectics*, pages 3–25.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of ndcg type ranking measures. In *Conference on learning theory*, pages 25–54. PMLR.

Yu Zhang and Qiang Yang. 2021. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609.